View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Toulouse Capitole Publications

SOCIAL PREFERENCES OR SACRED VALUES? THEORY AND EVIDENCE OF DEONTOLOGICAL MOTIVATIONS

Daniel L. Chen and Martin Schonger*

Abstract Recent advances in economic theory, largely motivated by experimental findings, have led to the adoption of models of human behavior where decision-makers not only take into consideration their own payoff but also others' payoffs and any potential consequences of these payoffs. Investigations of deontological motivations, where decision-makers make their choice not only based on the consequences of a decision but also the decision *per se* have been rare. We provide a formal interpretation of major moral philosophies and a revealed preference method to distinguish the *presence* of deontological motivations from a purely consequentialist decision-maker whose preferences satisfy first-order stochastic dominance.

JEL Codes: D63; D64; D91; K00

Keywords: Consequentialism, deontological motivations, normative commitments, social preferences, revealed preference, decision theory, random lottery incentive method

*Daniel L. Chen, daniel.chen@iast.fr, Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France; Martin Schonger, mschonger@ethz.ch, ETH Zurich, Law and Economics. First draft: May 2009. Current draft: August 2019. Latest version available at: http://users.nber.org/∼dlchen/papers/Social_Preferences_or_Sacred_Values.pdf. We thank research assistants and numerous colleagues at several universities and conferences. This project was conducted while Chen received funding from the Alfred P. Sloan Foundation (Grant No. 2018-11245), European Research Council (Grant No. 614708), Swiss National Science Foundation (Grant Nos. 100018-152678 and 106014-150820), Ewing Marion Kauffman Foundation, Institute for Humane Studies, John M. Olin Foundation, Agence Nationale de la Recherche, and Templeton Foundation (Grant No. 22420).

Your friend is hiding in your house from a murderer. The murderer arrives and asks you whether your friend is hiding in your house. Assuming you cannot stay silent, should you lie or tell the truth? (Kant 1797)

1. INTRODUCTION

There is a classic divide between the consequentialist view that optimal policy should be calculated from considerations of costs and benefits and an alternative view, held by many non-economists, that policy should be determined deontologically—people, society, and judges have duties; from duties, they derive what is the correct law, right, and just. This paper asks the behavioral question: Are there deontological motivations? If so, how would these motivations be formally modeled? What do deontological motivations imply for economics? What puzzles can be explained that elude standard models?

In the last few decades, economic theory has gradually expanded the domain of preferences. The homo-oeconomics view that individuals are only motivated by selfish material consequences confronted mounting evidence, usually in the lab, that individuals had other motivations–such as fairness (e.g. Rabin 1993), inequality aversion (e.g. Fehr and Schmidt 1999), reputation (e.g., McCabe et al. 2003; Falk and Fischbacher 2006; Dana et al. 2006, 2007), or social image (e.g., Bénabou and Tirole 2006; Andreoni and Bernheim 2009), to name a few. A common feature of these models is that motivations are consequentialist, in the sense that preferences are over acts because of their effects. These preferences are prominently characterized as *hypothetical imperatives*—preferences over acts because of their consequences—as opposed to *categorical imperatives*—preferences over acts regardless of their consequences—which Immanuel Kant (1797) called deontological motivations.

In general, the presence of deontological motivations is hard to detect. The usual method to measure deontological motivations is through survey or vignettes that present ethical dilemmas, like the moral trolley problem (Foot 1967). What our paper develops is a revealed preference method and a theorem that predicts invariance in the thought experiment if people are motivated solely under consequentialist motivations—but, if deontological motivations are *present, in combination* with consequentialist ones—then this thought experiment will reveal variance.

We can put an abstract form to the categorical imperative. Think of a decision-maker (DM) making a decision d. We want to separate the motivation for the decision from the motivation for its consequences. Consequences can be broad, including reputation, inequality, warm glow, and own

payoffs. Consequences x is a function of the state of nature and decision d. There are two states, in the consequential state, d becomes common knowledge and is implemented. In the non-consequential state, d remains unknown to anyone, including the experimenter. With consequentialism, preferences are over lotteries (Anscombe and Aumann 1963). With deontological motivations, d matters *per se*, even in the non-consequential state. To illustrate, Kant said in his axe-murderer hypothetical, "You must not lie," no matter what are the consequences.

Think of d^1, d^2, .., $d^{|D|}$, as possible decisions. Our experiment varies the probability that the decision is implemented. With some probability, π, your decision is implemented—has consequences—and with $1-\pi$, your decision has no consequence. So x^C is a function of the decision and x^N, some constant outcome that's invariant to your decision. This thought experiment can apply to any decision with a moral element, but we illustrate our theorem using the dictator game as it is one of the games most used in the academic literature.[1] In a dictator game, you have your endowment ω, and you can donate anywhere from 0 to ω. In our thought experiment, with some probability π, decisions are carried out. The recipient receives d and you receive the $\omega - d$. With probability $1-\pi$, your decision is not implemented—recipient receives κ and you keep the remainder. Subjects put their irrevocable decisions anonymously in sealed envelopes, and their envelope is *shredded* with some probability with a public randomization device and the probability is known in advance (Figure 1). Shredding means that the decision has no consequences, not even through the experimenter.[2] The decision only has consequences if the envelope is opened. Our shredding criterion for deontological motivations parallels Kant's discussion of his own thought experiment. Kant, likewise, allowed for uncertainty—the possibility that the decision has the ultimate adverse consequence or has no consequences[3]—but "to be truthful in all declarations is a sacred and unconditionally commanding law of reason that admits no expediency whatsoever." Kant's categorical imperative focused on the act itself rather than the expected consequences of an act. It is this motivation we seek to model and

[1] A google scholar search for "dictator game" yields 14 thousand articles, "trust game" 13 thousand, and "public goods game" 12 thousand. A search for "ultimatum game" yields roughly 22 thousand results and this is studied in more detail in Chen and Schonger (2015). We also considered "lying game" and "lying aversion", which only appear roughly 600 times each. "Prisoner's dilemma" appears roughly 49 thousand times, but we chose to focus on the simplest decision without strategic considerations. Andreoni and Bernheim (2009) likewise use the dictator game in their study of social image motivations. An analog can also be made between donations and tithing or tax morale.

[2] This eliminates motivations related to experimenter observation (Cilliers et al. 2015) and any altruism related to the societal good of providing one's data for science.

[3] "It is indeed possible that after you have honestly answered Yes to the murderer's question as to whether the intended victim is in the house, the latter went out unobserved and thus eluded the murderer, so that the deed would not have come about."

FIGURE 1.— Lab Implementation

uncover behaviorally.

The closest field analogs of our experiment may be found in two recent papers. First, Bergstrom et al. (2009) examined the decision to sign-up as a bone marrow donor. With some probability the decision to sign-up has consequences, such that the recipient receives bone marrow and the donor undergoes expensive and painful surgery. Bergstrom et al. (2009) found that those less likely to sign up to be a bone marrow donor came from ethnic groups that, due to genetic match and need, were more likely to be called off the list to donate. They argue this pattern to be a puzzle. Second, Choi et al. (2012) studied the decision not to abort a fetus with Down Syndrome. Prospective parents varied in the probability the decision to abort had consequences. They found that as the prospect became more real (hypothetical, high-risk, vs. diagnosed), parents were more likely to abort. In

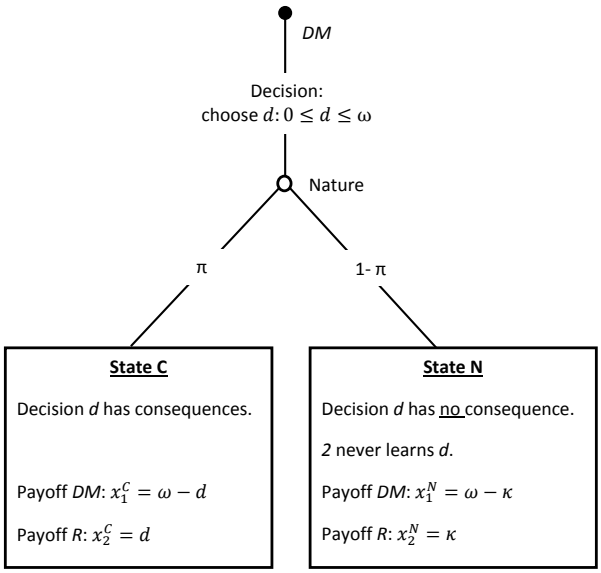
both Bergstrom et al. (2009) and Choi et al. (2012), as π decreased, people became more likely to choose a decision that might be interpreted as deontological. However, in both settings, d is not irrevocable and not anonymous and π is not exogenous, leaving room for potential confounders. In our laboratory setting, d is irrevocable and anonymous and π is exogenously assigned to the individual.

Formally, we show that pure deontologists following the categorical imperative would not change their behavior as the probability changes, but, counter-intuitively, it turns out that pure consequentialists also do not change their behavior. We provide a graphical and formal proof that someone who satisfies the behavioral assumption of first order stochastic dominance (FOSD) and is purely consequentialist will not change their behavior as the probability changes. Simply put, the decision-maker is choosing between lotteries G and F, so if G first-order stochastically dominates F with respect to $\succsim$ (i.e., if for all x': $\sum_{x:x'\succsim x} G(x) \leq \sum_{x:x'\succsim x} F(x)$), then if a decision d is optimal for one probability π, it is the optimal d for all probabilities. As a corollary, we can state the result with expected utility (a stronger behavioral assumption than first order stochastic dominance).[4] For the decision-maker donating the marginal penny, the marginal benefit of donating is the recipient's well-being and any social consequence of that increase. The marginal cost is to give up that penny. The decision-maker equates the marginal benefits and marginal costs. As the probability that the decision is implemented falls, then both the marginal benefits and costs fall equally, so the decision-maker still makes the same decision on the margin because the indirect objective function is proportional to the utility of the decision implemented with certainty.

To bridge our theorem to experimental evidence, our first study uses subjects in a lab. We asked subjects to choose an amount for a charitable recipient (as illustrated in Figure 2), a third-party aid organization. We found that subjects became 50% more charitable when the decision was hypothetical. Our second piece of evidence uses an online anonymous experiment, allowing large samples and very low implementation probabilities; but a difference is that d is observed by the experimenter even in the non-consequential state. If motives related to the experimenter or the study are strong, we may expect less variance. We found that subjects became 33% more charitable as the decision became hypothetical.

[4]The corollary holds since expected utility's independence axiom implies the axioms of first order stochastic dominance.

FIGURE 2.— Actual Experiment



It is possible that subjects become more charitable as the implementation probability falls because they value some kind of ex-ante fairness involving preferences over expected outcomes (Trautmann 2009; Krawczyk 2011; Chlaß et al. 2014).[5] While this is not a deontological motivation in Kant's typology, it is a behavioral motivation that can confound the interpretation of our results. To investigate that motive, the two experiments also had a treatment arm where the non-consequential state involves the entire sum being donated. Our data can rule out an expected-income targeter, who should have become *less* generous in response to reductions in π. Our data can also rule out other ex-ante fairness motivations. Finally, our data on decision time suggests that cognition costs are also not the explanation for variance between high and low π.

Our third piece of evidence illustrates how assumptions on the curvature of motives together with data on decision variance can inform how individuals trade-off between consequentialist and deontological motives. We use standard parameterizations of a structural model—consequentialist motivations are estimated with a classic Fehr-Schmidt inequity aversion utility while deontological motivations are estimated as a bliss point as in Cappelen et al. (2007) and Cappelen et al. (2013). The variation in our data generated by the experiment is consistent with largely deontological rather than consequentialist motives under the entire range of standard inequity aversion parameters.

[5] See also work on distributive justice (Elizabeth Hoffman 1985; Konow 2000) and procedural fairness (Gibson et al. 2013; Brock et al. 2013).

Like Bergstrom et al. (2009) observing more bone marrow donations and Choi et al. (2012) observing more decisions to not abort when the decisions were more hypothetical, we see d increases when π falls. What our model suggests is that as the probability falls, the (net negative) consequences of carrying out the act falls, but the (deontological) benefits of the act remain high. Moreover, the direction of change can give insight into the location of the maximand for an individual's duty (relative to the consequentialist maximand). Assuming the pure deontologist's maximand is higher than the pure consequentialist's maximand, reducing the probability results in decisions that are more deontological.

Our paper makes two contributions to the economic literature—theoretical and experimental. Economic models have thus far focused on *hypothetical imperatives* (preferences over acts because of their consequences). This interpretation is supported by Sobel's (2005) extensive literature review of interdependent preferences, part of which offered a typology of non-homo-oeconomics models. In one class are Chicago School models that model preferences over general commodities transformed into consumption goods. In another class are identity models (e.g., Akerlof and Kranton 2000) with utility functions over actions and an identity that incorporates the prescriptions that indicate the identity-appropriate behavior. Sobel noted that "the models of Akerlof–Kranton and Stigler–Becker are .. mathematically identical. It is curious that these formally equivalent approaches are associated with schools of thought that often are viewed as opposites. The theories are identical because they are consistent with precisely the same set of observations." In our reading, both classes of models fall under the hypothetical imperative: Chicago agents choose between *quantities*, but do not have preferences over *choices* vs. preferences over *quantities.* In identity models, agents choose *acts*, but do not have preferences over *acts* vs. preferences over *consequences of acts.* The categorical imperative would distinguish these preferences. Our thought experiment and shredding criterion likewise distinguishes choices from quantities and acts from consequences of acts.

Empirical researchers also have assumed that choices do not enter the utility function separate from the causal effects of choices. For example, in the random lottery incentive, experimental subjects make many choices, but only one of them is chosen at random to be implemented. In this oft-used method in experimental economics, if decisions involve a deontological element, the degree of pro-social behavior may be over-estimated, the lower the likelihood of implementation. In the strategy method—another method often used to increase statistical power—subjects make many

choices corresponding to possible states that may depend on what other subjects choose, but only a fraction of decisions count for pay. Deontological motives would imply that this bias from random lottery incentives would never disappear, no matter how high the stakes are.

Likewise, in surveys (which includes contingent valuation), subjects report preferences in non-consequentialist settings (e.g., valuation of an environmental good in a hypothetical scenario), and the decisions may change as the decision becomes more likely to be implemented.[6] In measuring willingness to pay, subjects report a price that is implemented if it is higher than a randomly generated price in the Becker-DeGroot-Marschak method. In the Vickrey auction, bidders submit written bids that are consequential if it is the highest bid. The higher the price, the more likely the decision has consequences. In market design data, subjects report preferences over choices over schools whose likelihood of being consequential varies.

Notably, our operationalization of deontological motives—choosing a decision regardless of the likelihood of implementation (i.e., irrespective of the consequences)—bears close similarity to the concept of legitimacy defined in psychology. Tyler (1997) characterized perceived legitimacy of laws and organizations as that which motivates obedience to rules irrespective of likelihood of reward or punishment. The remainder of the paper is organized as follows. Section 2 presents related literature. Section 3 defines consequentialism, deontologicalism, and mixed motivations. It proves that behavior is invariant to the probability for pure consequentialism or deontologicalism, but varies with mixed motives. Section 4 describes the empirical results. Section 5 concludes.

2. RELATED LITERATURE

Adam Smith's (1761) impartial spectator in The Theory of Moral Sentiments may have been deontological though perhaps also consequentialist.

> "The patriot who lays down his life for... this society, appears to act with the most exact propriety. He appears to view himself in the light in which the impartial spectator naturally and necessarily views him, ... bound at all times to sacrifice and devote himself to the safety, to the service, and even to the *glory of the greater* ... . But though this sacrifice appears to be perfectly just and proper, we know how difficult it is... and how few people are capable of making it." (emphasis added) (Smith 1761).

[6] Papers on the strategy method (Chen and Schonger 2015) and survey design (Cavaille et al. 2018) offer fuller, formal treatments of the issues, literature reviews, meta-analysis, and new experiments that complement the current study.

There is a vast economics literature on concepts related to deontological motivations. We refer the reader to Sobel's (2005) extensive literature review and focus our discussion here to subsequent work.[7]

The three closest theoretical developments may be as follows. First, deontological motivations may relate to identity investment. In Bénabou and Tirole (2011), moral decision-making is modeled as a form of identity investment that prevents future deviant behavior. Here, motives can be deontological or consequentialist. The DM cares about the fact that the decision is implemented. Second, deontological motivations may also relate to expressive motives. People may participate in elections even when their vote is not pivotal because of a perceived duty to vote (Riker and Ordeshook 1968). Feddersen et al. (2009) and Shayo and Harel (2012) formalize that insight where individuals obtain a small positive payoff by the act of voting for an option independent of the electoral outcome, which they test with experiments by varying the probability of being pivotal. Here, expressive motives can be deontological or consequentialist. The DM cares about the fact that the vote is cast. Election outcomes are public, so a message is sent to the public and vote share can affect the legitimacy of a candidate. DellaVigna et al. (2013) show experimentally that act of voting includes motives to tell others. Third, deontological motivations may also relate to "homo kantiensis", whose preferences are ones that are socially optimal when everyone else also holds that view (Alger and Weibull 2013). Alger and Weibull (2013) report that such preferences are selected for when preferences rather than strategies are the unit of selection and they find that preferences that are a convex combination of homo oeconomicus and homo kantiensis will be evolutionarily stable. Here, motives can be deontological or consequentialist. The DM cares about the outcome of everyone making the same decision.[8]

[7]Some exceptions not covered in the literature review include earlier work by Binmore (1994), arguing that John Rawls justifies the "original position" behind the veil of ignorance as an operationaliziation of Kant's categorical imperative, and by Harsanyi (1977), saying that empathetic preferences—requiring us to put ourselves in the position of another to see things from their point of view—is, under mild conditions, important for an implementation of Rawls' theory of justice.

[8]Warm glow motives can also be deontological or consequentialist. In an earlier theoretical contribution, Andreoni (1990) points out that DMs in a public goods contribution framework can derive utility not only from the total amount of the public good G provided, but also from her contribution g. However, the author suggests in Andreoni and Bernheim (2009) that social audience motivations can provide micro-foundation for the warm glow. Thus, the DM cares about the fact that the decision is observed. In other work, Ellingsen and Johannesson (2008) has a utility function incorporating own payoff, others' payoff, and how others think of me. The DM cares about the consequences of actions. Deontological motivations may also relate to guilt aversion (Battigalli and Dufwenberg 2007). The prototypical cause would be the infliction of harm or distress on the recipient, which can be deontological or consequentialist.

A large experimental literature has been interested in studying the motives for pro-social behavior. The shredding criterion can be distinguished from the experimental paradigm that varies the probability that one's decision will have an impact, since in those paradigms the decision-maker experiences the *cost* of helping in both states of the world (Batson et al. 1991; Smith et al. 1989).[9] In other experimental paradigms (Feddersen et al. 2009; Shayo and Harel 2012; Grossman 2015; Gneezy 2005), the decision-maker experiences the *benefits* of the decision in both states of the world. In a contemporaneous research design that is related, Andreoni and Bernheim (2009) use a modified dictator game with random implementation probabilities, but there are five differences. First, we make the recipient a charitable organization outside the lab; in their study, the recipients are in the room observing the decision and dictators become *more* generous as the probability of implementation increases because they are motivated by their social audience. Second, we make both the probability and the realization of the state of nature public; in their study, recipients observe the probability but not the fact that nature chose the outcome. Third, in their study, they acknowledge there may be motivations regarding what the experimenter infers and regard this as a confound; our lab experiment shreds decisions, which directly removes that confound. Fourth, their study uses the strategy method and subjects play several games, whereas in our study, each subject sees only one probability and we do not use the strategy method. Fifth, they recognize the importance of not using within-subject variation for any particular game; we directly remove sequence effects and contrast effects (for example, if an experimenter asks two questions with a higher and lower probability, subjects may feel the right answer is to give more in one scenario, which would be a confound for our invariance theorem).[10]

Large literatures outside of economics, such as psychology, political science, sociology, and law have discussed concepts related to deontological motives. Sacred values and taboos are also often interpreted as pertaining to duty, and that some actions cannot be evaluated through costs and benefits (Tetlock 2003). Some of these have been analyzed by economists—conflicts of sacred values (Bowles and Polania-Reyes 2012), repugnance (Roth 2007; Mankiw and Weinzierl 2010), and saving

[9]These studies examine whether one's help *actually* helps, rather than whether one's help *will be carried out*: the cost of the decision is experienced by subjects whether or not their decision to help is effective. These studies find, like Andreoni and Bernheim (2009), that as the probability falls, generosity declines, while we find the opposite.

[10]In another contemporaneous study, Grossman (2015) also uses a modified dictator game with random implementation probabilities, but each participant played the role of dictator and served as recipient for someone else. The study does not shred the decisions, so the decision's contribution is still a consequence. More broadly, we rule out motives related to the beliefs of others since the third-party aid organization is unaware of the subject.

the lives of mice (Falk and Szech 2013). Besley (2005) has argued to screen for deontological motivations in business leaders, politicians, or judges. In contrast, Kaplow and Shavell (2006) criticize relying on non-consequentialist motivations in optimal policy design as it would necessarily harm some individuals.

3. THEORY

In *The Stanford Encyclopedia of Philosophy*, Sinnott-Armstrong (2012) define consequentialism as, "the view that normative properties depend only on consequences" and explains that "[c]onsequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about." Utilitarianism is one example of a consequentialist moral philosophy (Bentham 1791); in fact any welfarist view is consequentialist (Arrow 2012). By contrast, deontological ethics holds that "some choices cannot be justified by their effects—that no matter how morally good their consequences, some choices are morally forbidden." (Alexander and Moore 2012).[11]

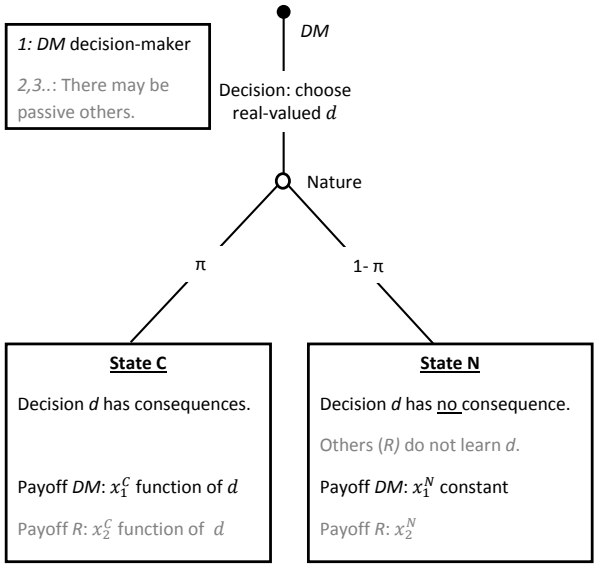
We introduce our thought experiment and focus on this definition of consequentialism and the invariance theorem first. We illustrate the intuition for the theorem under expected utility (this intuition is a corollary of the main theorem), a graphical proof of the invariance theorem, and then the formal statement of the assumptions along with the theorem itself. Next, we formalize deontological motivations as a lexicographic preference–duty first, then consequences–and show invariance still holds. We then show variance when individuals have both consequentialism and deontological motivations and the direction of change under additive separability.

3.1. *Thought Experiment*

The idea to identify non-consequentialist motivations by varying the probability of the DM's decision being consequential guides this paper. The DM has a real-valued choice variable d which influences both her own monetary payoff x_1 as well as the payoff x_2 of a recipient R. There are two states of the world, state C and state N. In state C, the DM's decision d fully determines both x_1 and x_2. In state N, both x_1 and x_2 take exogenously given values, and the decision d has no impact at all. Thus, in state C, the decision is consequential, while in state N, it is not. After DM chooses

[11] Virtues ethics, which originates in the work of Plato and Aristotle, would also be a non-consequentialist motivation we seek to uncover. To economize on terminology, we will only refer to deontological ethics. We also make no distinction between positive and negative duties.

FIGURE 3.— Thought Experiment: General Idea



d, nature randomly decides which state is realized. State C occurs with probability $\pi > 0$, state N with probability $1-\pi$. The structure of the game is public, but the decision d is only known to DM. In state N, therefore, R has no way of knowing d, but, in state C, R knows d, indeed he can infer it from x_2. Superscripts indicate the realized state, so that the payoffs are (x_1^C, x_2^C) in state C, and (x_1^N, x_2^N) in state N. Figure 3 illustrates this.

This general experimental design could be used for many morally relevant decisions; here we apply our identification method to the dictator game and thus to the moral decision to share. As shown in Figure 2, the DM receives an endowment of ω, and must decide how much to give to R. She may choose any d such that $0 \leqslant d \leqslant \omega$ and the resulting payoffs are $x_1^C = \omega - d$ and $x_2^C = d$. For $\pi = 1$, the game thus reduces to the standard dictator game. In state N, a pre-determined, exogenous κ will be implemented, where $0 \leqslant \kappa \leqslant \omega$, and $x_1^N = \omega - \kappa$ and $x_2^N = \kappa$ are the resulting payoffs.

3.2. *Intuition*

We illustrate the intuition of the invariance theorem under expected utility. Given expected utility, the DM maximizes:

$$E[u(x,d)] = \pi u(x_1^C, x_2^C, d) + (1-\pi)u(x_1^N, x_2^N, d)$$

and her indirect objective function in case of the dictator game can be written as:

$$V(d) = \pi u(\omega - d, d, d) + (1 - \pi)u(\omega - \kappa, \kappa, d).$$

Limiting attention to pure consequentialists, the problem simplifies to:

$$E[u(x)] = \pi u(x_1^C, x_2^C) + (1 - \pi)u(x_1^N, x_2^N)$$

and the indirect objective function to:

$$V(d) = \pi u(\omega - d, d) + (1 - \pi)u(\omega - \kappa, \kappa).$$

Note that now the d does not enter in the second term, which corresponds to state N. The indirect objective function is proportional to $u(\omega - d, d)$, so $\frac{\partial d^*}{\partial \pi} = 0$.

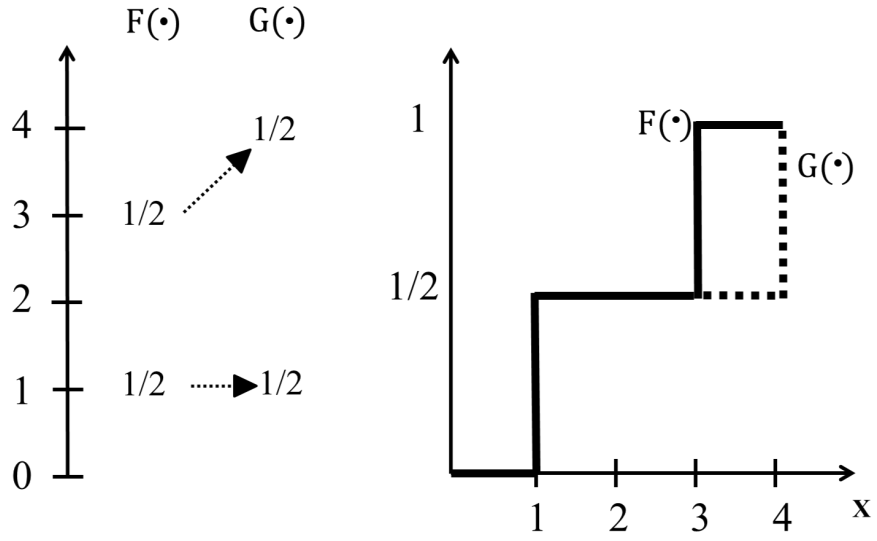
3.3. *Graphical proof*

In the previous subsection, we have seen that if the DM satisfies the axioms of expected utility then if d^* is not constant in the probability she cannot be a consequentialist. Put differently, if we observe a DM to vary her decision in the probability we would reject the joint hypothesis that the DM is a consequentialist and an expected-utility maximizer. Since expected utility theory often fails to describe behavior (Starmer 2000) such a joint test would tell us little about whether consequentialism or expected utility or both were rejected. It is therefore desirable to have much weaker assumptions about decision-making under objective uncertainty than expected utility theory. Here we show that first-order stochastic dominance is sufficient for the result.

First, we provide a graphical sketch of the invariance proof. That is, someone who satisfies the behavioral assumption of preference relations of FOSD and is purely consequentialist will not change their behavior as the probability changes. The left-hand side of Figure 4 provides an example of FOSD. Think of an ordering over outcomes, 0, 1, 2, 3, and 4 on the Y-axis and the corresponding lotteries F and G. G looks better than F since instead of getting 3, sometimes the DM gets 4. Formally, G first-order stochastically dominates F with respect to $\succsim$ if for all x': $\sum_{x:x' \succsim x} G(x) \leq \sum_{x:x' \succsim x} F(x)$.

For every outcome x', the probability of any outcome worse than x' is lower under G than under F. That can be represented graphically on the right, as a CDF. For the proof, recall that decisions

FIGURE 4.— First Order Stochastic Dominance



are choices over lotteries like F and G. Suppose 1 is the non-consequentialist outcome, and let 3 or 4 be the active choice. What does changing the probability do? It moves the horizontal bar up and down. But G always FOSD F. So if a choice is optimal for one probability, it is the optimal choice for all probabilities.

3.4. *Formal statement of assumptions and theorem*

In our delineation, we try to adapt major concepts of moral philosophy to economics, and bring the precision of economic methodology, in particular revealed preference, to moral philosophy. It may seem odd to model deontological motivations by utility functions since one may view "utility" as a consequence, but since ours is a revealed preference approach, we follow the usual economics approach (Friedman and Savage 1948) of modeling decision-makers' behavior as if they maximized that objective function and refrain from interpreting the function as standing for utility or happiness.

We allow the utility u of the DM to be a function of her own monetary payoff x_1, as well as the monetary payoff of the recipient x_2 to capture consequentialist other-regarding motives, and d to capture deontological motives. In the general case with all motivations present, the Bernoulli utility function satisfies $u = u(x_1, x_2, d)$. The standard theories of decision-making by Savage (1972) and Anscombe and Aumann (1963) rely on the assumption that the domain of consequences is state independent.

DEFINITION 1 CONSEQUENTIALIST PREFERENCES: A preference is *consequentialist* if there exists a utility representation u such that $u = u(x)$.

We call a preference consequentialist-deontological if it incorporates concerns beyond the consequences, and considers actions or decisions that are good or bad per se:

DEFINITION 2 CONSEQUENTIALIST-DEONTOLOGICAL PREFERENCES: A preference is *consequentialist-deontological* if there exists a utility representation u such that $u = u(x, d)$.

Now let us turn to purely deontological preferences. At first, one might think they are simply mirroring the other extreme of consequentialist preferences and could thus be represented by $u = u(d)$. But, since duty is like an internal moral constraint, even fully satisfying one's duty may leave the DM with many morally permissible options rather than one unique choice. A deontologist can be formalized as having a lexicographic preference on decisions d and outcome x, with deontological before consequentialist motivations.

DEFINITION 3 DEONTOLOGICAL PREFERENCES: A preference is called *deontological* if there exist u, f such that $u = u(d)$, and $f = f(x)$, and for all $(x, d), (x', d')$: $(x, d) \succsim (x', d')$ if and only if $u(d) > u(d')$ or $[u(d) = u(d')$ and $f(x) \geqq f(x')]$.

It is possible to model purely deontological people as having a different choice set (Nozick 1974). But traditionally a choice set is the objective, external constraints facing a person and we call the internal constraints preferences. Thus, we model deontological moral constraints on the decision-maker as internal constraints, that is, as the first part of preferences in a lexicographic framework. The reason we do not model duty like a budget constraint but as part of preferences, and thus lexicographic is twofold: First, unlike budget constraints, internal moral constraints are not directly observable; second, for consequentialist-deontological preferences that feature a tradeoff rather than a lexicographic ordering of these motivations, one could not model duty as an inviolable constraint. This can be formalized as a lexicographic preference, with deontological before consequentialist motivations. Note that while economists may think of our method as detecting where a DM feels most duty among competing duties (i.e., the optimand of one's *greatest* duty rather than the optimand of

one's duty), some philosophers believe there is no possibility of a genuine conflict of duties in deontological ethical theory, which can distinguish between a duty-all-other-things-being-equal (prima facie duty) and a duty-all-things-considered (categorical duty) (Alexander and Moore 2012).

We delineate assumptions that allows us to experimentally identify with observable choice behavior whether subjects have preferences where both motivations are present (i.e., whether their preferences belong to the category of consequentialist-deontological preferences). The standard consequentialist approach to (and a central assumption for) choice under uncertainty is first-order stochastic dominance (FOSD). A wide variety of models of choice under uncertainty satisfies FOSD and thus falls within this framework, among them most prominently, expected utility theory, its generalization by Machina (1982), but also cumulative prospect theory (Tversky and Kahneman 1992) or rank-dependent utility theory (Quiggin 1982).

Following the canonical framework as laid out in Kreps (1988), let there be outcomes x. x can be a real valued vector. In the thought experiment, it would be $x = (x_1, x_2)$. Let the set of all x be finite and denote it by X. A probability measure on X is a function $p : X \to [0, 1]$ such that $\sum_{x \in X} p(x) = 1$. Let P be the set of all probability measures on X, and therefore, in the thought experiment, a subset of it, is the choice set of the decision-maker.

AXIOM 1 *(preference relation) Let $\succsim$ be a complete and transitive preference on P.*

Axiom 1 is the standard one saying that the preference relation is a complete ordering. It implicitly includes consequentialism since the preference relation is on P, that is, over lotteries that are over consequences x.

Next we define first-order stochastic dominance (FOSD). Often, definitions of FOSD are suitable only for preference relations that are monotonic in the real numbers, for example see Levhari et al. (1975). These definitions define FOSD with respect to the ordering induced by the real numbers, assuming that prices are vectors. It is important to define FOSD with respect to ordering over outcomes rather than the outcomes themselves.[12]

DEFINITION (FOSD) p first-order stochastically dominates q with respect to the ordering induced by $\succsim$, if for all x': $\sum_{x:x' \succsim x} p(x) \leq \sum_{x:x' \succsim x} q(x)$.

[12]FOSD over outcomes is inappropriate in the context of social preferences, which are often not monotonic due to envy or fairness concerns.

AXIOM *(FOSD) If p FOSD q with respect to the ordering induced by $\succsim$, then $p \succsim q$.*

DEFINITION (Strict FOSD) p strictly first-order stochastically dominates q with respect to the ordering induced by $\succsim$ if p FOSD q with respect to that ordering, and there exists an x' such that: $\sum_{x:x' \succsim x} p(x) < \sum_{x:x' \succsim x} q(x)$.

Formally, our theorem needs both strict FOSD and weak FOSD since strict FOSD does not imply weak FOSD.

AXIOM *(Strict FOSD) If p strictly FOSD q with respect to the ordering induced by $\succsim$, then $p \succ q$.*

The following theorem implies that in our thought experiment, changing the probability of being consequential π does not change the decision.

THEOREM 1 *If the DM satisfies the axioms Preference Relation, FOSD, and Strict FOSD, and there exist $x, x', x'' \in X'$ and $\pi \epsilon (0;1]$ such that $\pi x + (1-\pi)x'' \succcurlyeq \pi x' + (1-\pi)x''$, then for all $\pi' \epsilon (0;1] : \pi' x + (1-\pi')x'' \succcurlyeq \pi' x' + (1-\pi')x''$.*

It is this prediction of the theory that we will test and interpret a rejection of the prediction as evidence that people are not purely consequentialist. Proofs and additional theoretical discussion are relegated to Appendix A.

FACT 1 ***(Deontological preferences)*** *For purely deontological preferences the optimal decision d^* is constant in the probability π.*

This is because in these lexicographic preferences, a person is either pure deontological or pure consequentialist in comparing possible decisions. Formally, there is no trade-off. A lexicographic deontologist maximizes $u(d)$ first, then there is a compact set where she maximizes $v(x)$ next. Our theorem applies to either the pure consequentialist portion $v(x)$ or the deontological portion $u(d)$.

3.5. *Consequentialist-deontological preferences*

Next, we illustrate consequentialist-deontological preferences where the optimal decision changes as the probability of being consequentialist changes. For exposition, we do so in the context of Figure 2 and simplify notation such that the net consequences are a function of x_1.

EXAMPLE 1 $u = u(x_1, d) = x_1 + b(d)$, where $b_1 > 0$ and $b_{11} < 0$.

Then $V(d) = \pi(\omega - d) + (1 - \pi)(\omega - \kappa) + b(d)$ is strictly concave in d. The first-order condition is $b_1(d) = \pi$ and thus for an interior solution $\frac{\partial d^*}{\partial \pi} = \frac{1}{b_{11}(d)} < 0$. The second order condition is $b_{11}(d) < 0$. Note that if the consequentialist and deontological choice is the same, then the choice is still invariant to the implementation probability: $f_1(\omega - d) = b_1(d) = 0$, then $\frac{\partial d^*}{\partial \pi} = 0$.

For a slightly more general example: let $u(x_1, d) = f(x_1) + b(d)$. Then, $U(x_1, d) = \pi(f(x_1^C) + b(d)) + (1 - \pi)(f(x_1^N) + b(d))$ and $V(d) = \pi f(\omega - d) + (1 - \pi)f(\omega - \kappa) + b(d)$. The first order condition is: $\frac{\partial V(d)}{\partial d} = -\pi f_1(\omega - d) + b_1(d) = 0$. For d^* to be a maximum, the second order condition yields: $\frac{\partial^2 V(d)}{\partial d^2} = \pi f_{11}(\omega - d) + b_{11}(d) < 0$. Applying the implicit function theorem to the first order condition yields: $\frac{\partial d^*}{\partial \pi} = \frac{f_1(\omega - d^*)}{\pi f_{11}(\omega - d^*) + b_{11}(d^*)} < 0$, since utility is increasing in its own outcomes and the denominator which is the second derivative of the indirect objective function is negative. Note that the recipient's payoff is a function of the DM's payoffs, but as long as other-regarding concerns are concave then the sum of utility from its own payoffs and utility from others' payoffs is still concave and the above result holds. Decisions do not have to be continuous to obtain this result. If decisions are discrete, then the behavior of a mixed consequentialist-deontological person is jumpy (i.e., it weakly increases as her decision becomes less consequential).

For more complicated utility functions, non-additive or non-globally convex ones, it is possible to generate examples where $\frac{\partial d^*}{\partial \pi} = \frac{1}{b_{11}(d)} > 0$. Suppose the DM has preferences represented by $u = u(x_1, d)$. Assume that the first derivatives are positive (monotonicity), and that $u_{11} < 0$ and $u_{22} < 0$ (risk-aversion). Then the DM maximizes $V(d) = \pi u(\omega - d, d) + (1 - \pi)u(\omega - \kappa, d)$. The first order condition is $-\pi u_1(\omega - d, d) + \pi u_2(\omega - d, d) + (1 - \pi)u_2(\omega - \kappa, d) = 0$. By the implicit function theorem, and simplifying using the first order condition gives:

$$\frac{\partial d^*}{\partial \pi} = \frac{1}{\pi^2}[-2u_{12}(\omega - d, d) + u_{11}(\omega - d, d) + u_{22}(\omega - d, d) + \frac{1 - \pi}{\pi}u_{22}(\omega - \kappa, d)]^{-1}u_2(\omega - \kappa, d)$$

So for sufficiently negative $u_{12}(\omega - d, d)$ we can get $\frac{\partial d^*}{\partial \pi} > 0$. Utility functions that are not globally convex can lead to local maxima that, when the decision is less consequential, can lead to jumps to maxima involving lower d.[13]

[13] In related work, Chen et al. (2015) explores modeling and testing the shape of the cost of taking actions that one disagrees with morally or politically.

3.6. *Potential Confounds*

3.6.1. *Ex Ante Fairness*

A potential confound to testing the invariance theorem in an experiment is that people could have preferences over the lotteries themselves if they view them as procedures, rather than if their preferences are fundamentally driven by the prizes (consequences or the decision). In our experimental setup, for example a subject might target the expected income of the recipient, and thus vary the decision in the probability. This section shows formally that by varying κ we can test whether people have these ex-ante considerations. Targeting the recipient's expected income can be assessed by our research design by seeing if the sign of $\frac{\partial d^*}{\partial \pi}$ flips in the two treatment arms, one where κ is set at 0 and another where κ is set at the maximum.

EXAMPLE 2 Targeting the recipient's expected income. Consider the following preferences $U(x_1, x_2) = E[x_1] + a(E[x_2]) = \pi x_1^C + (1-\pi)x_1^N + a\left(\pi x_2^C + (1-\pi)x_2^N\right)$. Let a be a function that captures altruism and let it be strictly increasing and strictly concave. Note that this objective function is not linear in the probabilities. The indirect objective function is $V(d) = \pi(\omega - d) + (1-\pi)(\omega - \kappa) + a(\pi d + (1-\pi)\kappa)$. The first-order condition is $a_1(\pi d + (1-\pi)\kappa) = 1$. By the implicit function theorem, $\frac{\partial d^*}{\partial \pi} = \frac{\kappa - d^*}{\pi}$. Thus the optimal decision changes in the probability. In two special cases, it is easy to determine the sign of the derivative, even if d^* itself is not (yet) known: if $\kappa = 0$, then $\frac{\partial d^*}{\partial \pi} \leqq 0$, and if $\kappa = \omega$, then $\frac{\partial d^*}{\partial \pi} \geqq 0$.

Let us look at a more general case: $U = f(E[u(x_1)], E[\widetilde{u}(x_2)])$, where f is $f_1, f_2 > 0$ (strictly increasing), $f_{12}f_1f_2 - f_{11}f_2^2 - f_{22}f_1^2 > 0$ (strictly quasi-concave), ($f_{12}f_2 - f_{22}f_1 > 0$ and $f_{12}f_1 - f_{11}f_2 \geq 0$) or ($f_{12}f_2 - f_{22}f_1 \geq 0$ and $f_{12}f_1 - f_{11}f_2 > 0$) (strictly normal in in one argument, weakly normal in the other), $u, \widetilde{u}$ is $u_1, \widetilde{u}_1 > 0$ (strictly increasing), $u_{11}, \widetilde{u}_{11} \leq 0$ (weakly concave) and $\pi > 0$. Then, the indirect objective function is

$$V(d) = f\left(\pi u(\omega - d) + (1-\pi)u(\omega - \kappa), \pi\widetilde{u}(d) + (1-\pi)\widetilde{u}(\kappa)\right)$$

Note that $V(d)$ is globally strongly concave:

$$\frac{1}{\pi}\frac{\partial^2 V(d)}{(\partial d)^2} = -\left(2f_{12}f_1f_2 - f_{11}f_2^2 - f_{22}f_1^2\right)\frac{1}{f_2^2}\pi u_1^2(\omega - d) + f_1 u_{11}(\omega - d) + f_2\widetilde{u}_{11}(d) < 0$$

So, there exists a unique solution. The First-order condition for this problem is $\frac{\widetilde{u}_1(d)}{u_1(\omega-d)} - \frac{f_1}{f_2} = 0 \equiv F$. The FOC defines d^* implicitly as a function of π. By the implicit function theorem $\frac{\partial d^*}{\partial \pi} = -\frac{\frac{\partial F(d^*,\pi)}{\partial \pi}}{\frac{\partial F(d^*,\pi)}{\partial d^*}}$. As $\frac{\partial F(d^*,\pi)}{\partial d^*}$ has sign of $\frac{\partial^2 V(d)}{(\partial d)^2} < 0$: $sgn\left(\frac{\partial d^*}{\partial \pi}\right) = sgn\left(\frac{\partial F(d^*,\pi)}{\partial \pi}\right)$. It can be shown that:

$$\frac{\partial F(d^*,\pi)}{\partial \pi} = \frac{\widetilde{u}_1(d^*)}{f_1}(f_{12}f_1 - f_{11}f_2)\left[u(\omega - d^*) - u(\omega - \kappa)\right] + \frac{u_1(\omega - d^*)}{f_2}(f_{12}f_2 - f_{22}f_1)\left[\widetilde{u}(\kappa) - \widetilde{u}(d^*)\right]$$

So the sign of $\frac{\partial d^*}{\partial \pi}(\pi)$ depends on the difference between $d^*(\pi)$ and κ:

For $d^*(\pi) = \kappa$: $\frac{\partial F(d^*,\pi)}{\partial \pi} = 0$ thus $\frac{\partial d^*}{\partial \pi}(\pi) = 0$

For $d^*(\pi) < \kappa$: $\frac{\partial F(d^*,\pi)}{\partial \pi} > 0$ thus $\frac{\partial d^*}{\partial \pi}(\pi) > 0$

For $d^*(\pi) > \kappa$: $\frac{\partial F(d^*,\pi)}{\partial \pi} < 0$ thus $\frac{\partial d^*}{\partial \pi}(\pi) < 0$

Now if $\kappa = 0$, then $\frac{\partial d^*}{\partial \pi} \leqslant 0$, while for $\kappa = \omega$, $\frac{\partial d^*}{\partial \pi} \geqslant 0$.

Thus experimentally, by varying κ we can test whether people have these ex-ante considerations. In sum, targeting the recipient's expected income can be assessed by our research design by seeing if the sign of $\frac{\partial d^*}{\partial \pi}$ flips in the two treatment arms. Motivations pertaining to forms of residual uncertainty that take into account ex-ante considerations but mix them with ex-post considerations would also predict the sign to flip.

3.6.2. *Cognition Costs*

Another explanation for variance in the probability might be cognition costs. Cognition costs are a consequence, but unlike the other consequences, they are not captured in our consequentialist framework since they are incurred during the decision and are a consequence that even arises if the non-consequential state is realized. Formal modeling and experimental test of cognition costs seems to be rare in the literature. For a previous example, albeit one that does not have the decision-maker solve the metaproblem optimally, see Wilcox (1993). This section shows that a cognition-costs model

would predict that 1) *time spent on the survey also changes with* π as d changes. Our research design also provides a second test: 2) *Subjects with greater cognition costs should have* $\frac{\delta d}{\delta \pi} = 0$ *for a larger range of* π near 0.

To fix ideas, consider the following model: $u = u(x_1, x_2, \gamma)$, where $u_1, u_2 > 0$, $u_\gamma < 0$ and $\gamma \geq 0$. In addition, let us assume that utility is continuous. The DM can compute the optimal decision, but to do so, she incurs a cognition cost $\gamma > 0$, otherwise she can make a heuristic (fixed) choice $\bar{d}$ for which (normalized) costs are 0. We have no model of what the heuristic choice is, and in principle it could be anything. Suppose the heuristic choice tends to be a cooperative or fair one (Rand et al. 2012) so, for example, the reader might think of $\bar{d} = \frac{\omega}{2}$. In any case, expected utility from the heuristic choice is $V(\bar{d}) = \pi u(\omega - \bar{d}, \bar{d}, 0) + (1 - \pi)u(\omega - \kappa, \kappa, 0)$. By contrast, for a non-heuristic choice, $V(d) = \pi u(\omega - d, d, \gamma) + (1 - \pi)u(\omega - \kappa, \kappa, \gamma)$. Define $\check{d} \equiv argmax V(d)$. Obviously, $\check{d}$ does not vary in π. The DM will choose to act heuristically if $V(\check{d}) < V(\bar{d})$ or

$$F(\pi) \equiv V(\check{d}) - V(\bar{d}) = \pi \left(u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0)\right) + (1 - \pi) \left(u(\omega - \kappa, \kappa, \gamma) - u(\omega - \kappa, \kappa, 0)\right) < 0$$

Since $(1 - \pi) \left(u(\omega - \kappa, \kappa, \gamma) - u(\omega - \kappa, \kappa, 0)\right) < 0$, we can distinguish two cases:

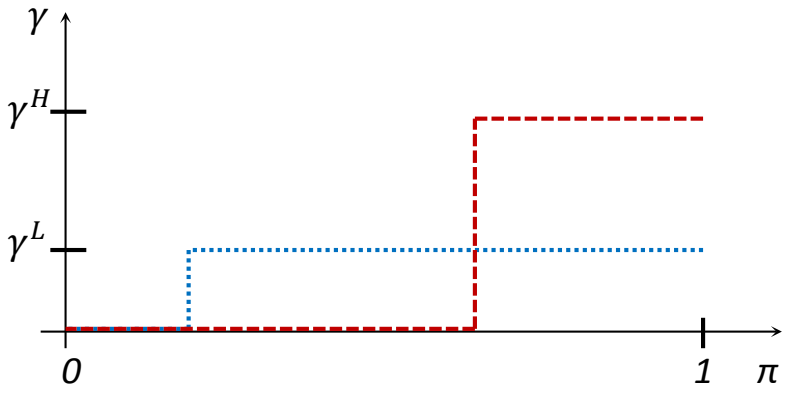
i) If $u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0) < 0$, $F(\pi)$ is always negative, so the person uses the heuristic choice, independent of π.

ii) In the other case, $u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0) > 0$, there exists a unique $\widetilde{\pi}$ with $0 < \widetilde{\pi} < 1$ such that $F(\widetilde{\pi}) = 0$, the person switches from heuristic to non heuristic. This derives from the fact that in this case $F(\pi)$ is strictly monotone in π, $F(0) < 0$ and $F(1) > 0$, so for probabilities of being consequential close to 1 computing is better, and for probabilities close to zero, the heuristic is better. Since $\check{d} \neq \bar{d}$, this means that such cognition costs predict that even a consequentialist DM will not be invariant to the probability. For the rest of this section, we will focus on this case.

Now suppose we vary the cognition cost, that is, we do an exercise in comparative statics and investigate how $\widetilde{\pi}$ varies in γ, and note that

$$\frac{\partial \widetilde{\pi}}{\partial \gamma} = \frac{-\widetilde{\pi} u_3(\omega - \check{d}, \check{d}, \gamma) - (1 - \widetilde{\pi}) u_3(\omega - \kappa, \kappa, \gamma)}{u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0) + u(\omega - \kappa, \kappa, 0) - u(\omega - \kappa, \kappa, \gamma)} > 0,$$

FIGURE 5.— S-Shape Cognition Costs



that is, the higher the cognition costs, the higher the threshold for probability being consequential such that computation is the better choice. Obviously, there are some very low γ and some very high γ such that locally, $\widetilde{\pi}$ is a constant function of γ, but there, the above assumptions are violated. Figure 5 shows when, as a function of a probability, someone would incur a given cognition cost. So if we could experimentally vary not only probability but also cognition costs and then observe it, the cognition cost story predicts the pattern shown in the figure.

In summary, variation in the decision d with respect to π is consistent with decision-makers switching to a heuristic $\bar{d}$, which may be higher or lower than the preferred choice $\check{d}$, leading to the inability to infer consequentialist-deontological preferences. If decision-makers have different γ or different $\bar{d}$, then we might observe a smooth $\frac{\delta d}{\delta \pi}$. A cognition-costs model, however, would predict that 1) *time spent on the survey also changes with* π as d changes. We also provide a second test: 2) *Subjects with greater cognition costs should have* $\frac{\delta d}{\delta \pi} = 0$ *for a larger range of* π near 0. An S-shape curve in cognition costs incurred and thus in decisions with respect to π, is more shifted, the higher cognition costs are. Figure 5 plots the cognition costs incurred against π. The dotted line is for the subject experiencing low cognition costs while the dashed line is for the subject experiencing high cognition costs.

3.6.3. *Self Image*

A conceptual distinction can be made between self-image and duty. Firstly, rather than self-image motives, "the central insight that gives deontology its name is that in moral reflection, the self discovers that an act ought to be done; it owes it to itself to do justice to this obligation" (Junker-

Kenny 2013). Secondly, self-image motives affect decisions simply when subjects anticipate finding out about peers (Bigenho and Martinez 2019). Individuals may also punish others who threaten their ego (Chen and Prescott 2016). Thirdly, self-signaling often modeled as an investment with long-term *consequences* (Bénabou and Tirole 2011).[14] To be sure, even purely deontological preferences likely have some neurobiological consequence, which perhaps could be studied with fMRIs.

4. RESULTS

4.1. *Lab Experiment*

We ran the lab experiment in Zurich using zTree (Fischbacher 2007). We asked subjects aged 18-30 to make a donation decision out of an endowment of 20Chf with the knowledge that we would shred their decision when it was not implemented. One session collected data from a classroom, but the procedures were the same and the endowment was 10Chf. All our results are reported in terms of percent donation. The donation recipient was Doctors Without Borders as we believed this organization to be more salient in German-speaking countries.

Participants first saw a demonstration of a public randomization device (Appendix B includes pictures and instructional materials) and a paper shredder; the shredding bin was opened to publicly verify that materials were truly going to be destroyed. Prior to the experiment, subjects were asked three IQ tasks. If at least one answer was correct, they proceeded to the donation decision and received information about their probability of implementation. We had a $2x2$ design: Subjects were randomly assigned to low ($\pi = \frac{3}{16}$) or high probability ($\pi = \frac{15}{16}$) of implementation, and to minimum ($\kappa = 0$) or maximum ($\kappa = \omega$) donation in the non-consequential state. The randomization wheel had sixteen numbers. We only mentioned one or three of these numbers to the subject depending on their π. The numbers between 1 and 16 were randomly chosen to minimize the potential influence of anchoring on the results. They were then asked to write a decision to be placed in a sealed envelope.

After the wheel was spun, envelopes that were to be destroyed were collected and shredded. The remainder were opened and participants were paid. Among 264 subjects, 71 envelopes were opened. We over-sampled subjects who received low probabilities. If we assign the same number of subjects to each treatment condition, far fewer data will be collected for $\pi = \frac{3}{16}$ treatment condition where

[14]This suggests that memory would be a mediating mechanism for self-image concerns. Thought experiments proposed by philosophers motivate an experimental design for future research where a decision-maker forgets (and knows that forgetting will occur), e.g., consider an individual taking an action knowing that he resides in Plato's cave.

only few envelopes are opened. We sought a roughly 1:1 ratio for the opened envelopes in the high and low π conditions. All results only analyze the decisions of envelopes opened as we do not have data for envelopes that were shredded.

Participants donated an average amount of 25% when π was high and 38% when π was low. Figure 6 disaggregates the results by κ and the vertical lines indicate means for each treatment group. Ex-ante fairness concerns would predict the effect of π to flip depending on the location of κ, but we observed an increase in donations (of roughly 50%) for both $\kappa = 0$ and $\kappa = Max$ treatments.

FIGURE 6.— Donation and π: Disaggregated by κ

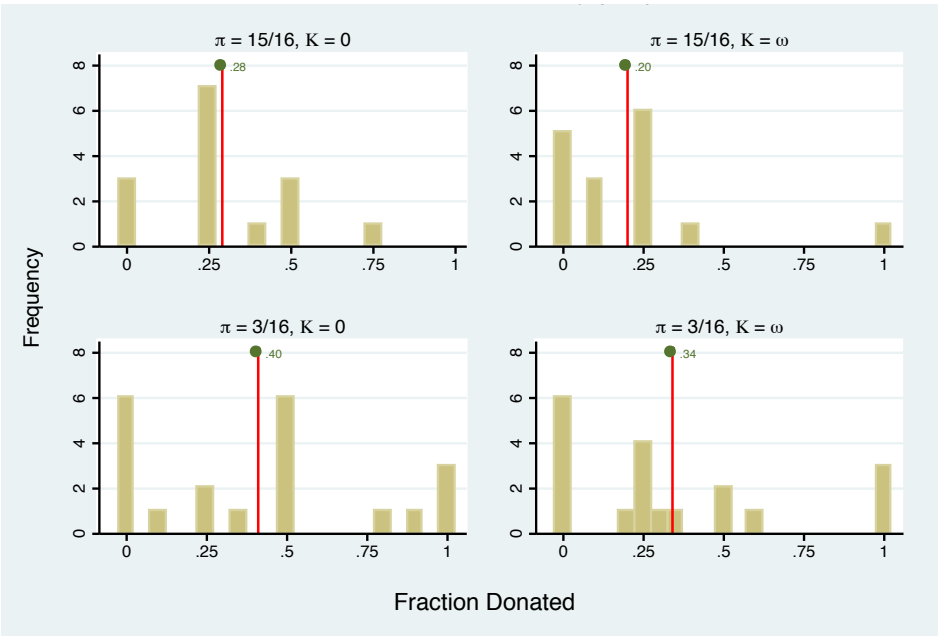


Table I reports regression results indicating that the change in donations is significant at the 10% level without κ fixed effects (Columns 1) or with κ fixed effects (Column 2). The estimates are stable. The R-square is 0.045 only including π. The magnitude of the effect is equivalent to roughly half the mean donation. Extrapolating linearly suggests that increasing the likelihood of implementation from 0% to 100% reduces the donation by roughly 17 percentage points. Columns 3-6 test for ex ante consequentialism. Increasing the likelihood of implementation from 0 to 1 strongly reduces the expected income by the donee (Columns 3-4) and strongly increases the expected giving of the donor (Columns 5-6), whether or not κ fixed effects are included. These effects are significant at the 1% level. The following presents additional visualizations of these results.

TABLE I

DONATION AND π: LINEAR REGRESSION

	Ordinary Least Squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected Income E(x_2)		Expected Giving (πd^*)	
Mean dep. var.	0.30		0.39		0.12	
% Consequential (π)	-0.176*	-0.159*	-0.259**	-0.278***	0.212***	0.219***
	(0.0978)	(0.0855)	(0.108)	(0.0802)	(0.0484)	(0.0452)
K Fixed Effects	N	Y	N	Y	N	Y
Observations	71	71	71	71	71	71
R-squared	0.045	0.292	0.077	0.506	0.218	0.339

Notes: Standard errors in parentheses. Raw data shown in Figures 4 and 5. * p < 0.10, ** p < 0.05, *** p < 0.01

Figure 7 graphically examines the ex ante fairness explanation. It shows that as π changes, expected income of the recipient is not fixed; it increases when κ is high and decreases when κ is low. When we calculate the expected income of a beneficiary, we use the data for subjects whose envelopes were opened and combine it probabilistically with κ.

FIGURE 7.— Expected Income E(x_2) and π: Disaggregated by κ

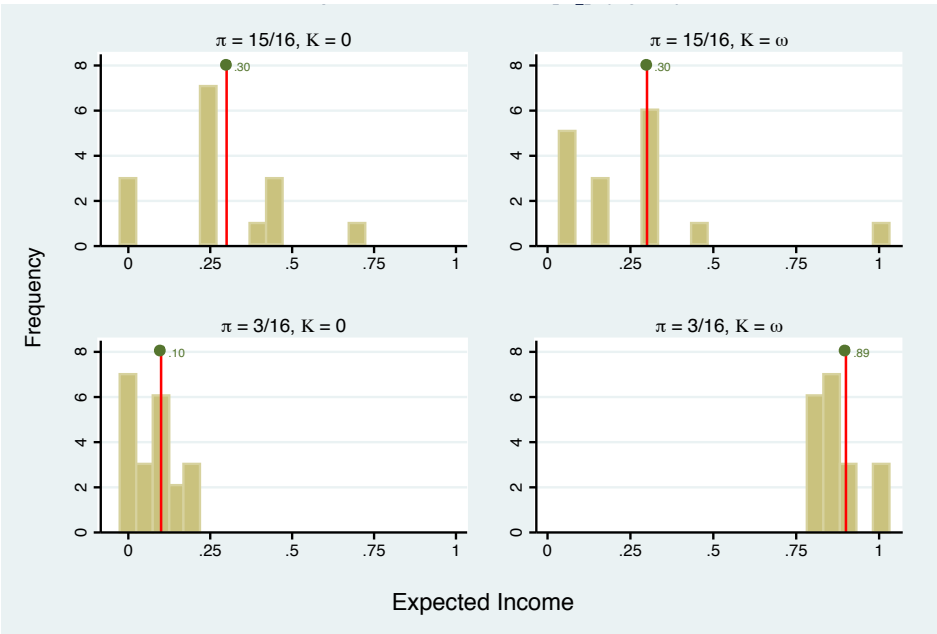


Figure 8 shows that as π changes, expected giving by the decision-maker is also not fixed. Expected giving does not depend on κ. It only depends on d and π. Our results indicate that for both κ, expected giving drops by two-thirds as π goes from high to low. The statistical significance (1% level) of the mean impact is displayed in Columns 5 and 6 of Table I.

FIGURE 8.— Expected Giving (πd^*) and π: Disaggregated by κ

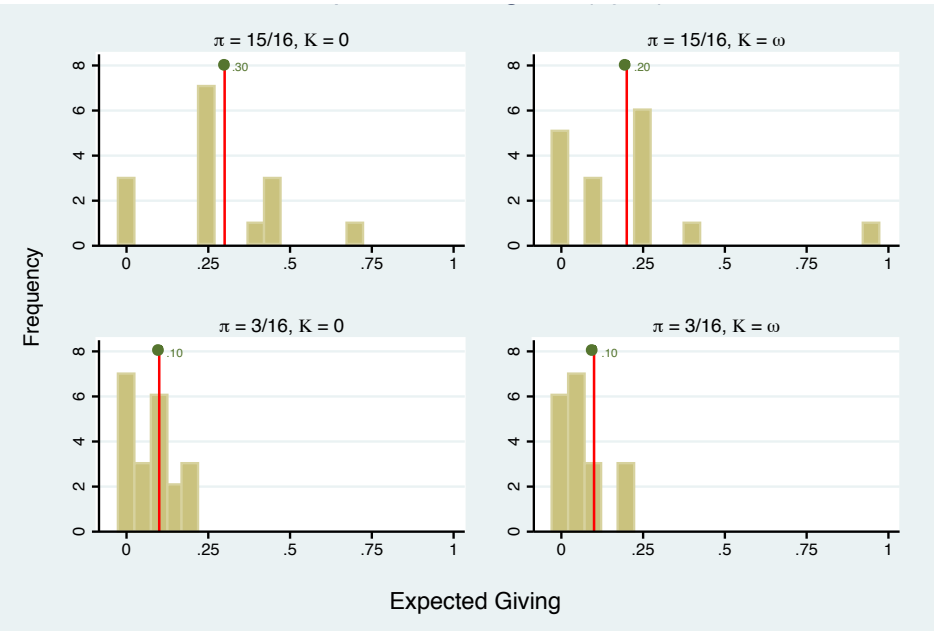


Table II presents Mood's median tests of the null hypothesis that medians of the two populations are identical. It has low power relative to the Mann Whitney test, but is preferred when the variance is not equal in different groups. We can see the variances are different in Figure 6. The median tests report significant differences at the 5% level for π and for κ.

TABLE II

DONATION AND π: NON-PARAMETRIC TESTS

	Non-parametric test for equality of medians, 2-sided test (p-values)
Thresholds	Pooled
$\pi = 3/16$ vs. $\pi = 15/16$	0.04
K = 0 vs. K = Max	0.01

4.2. *Online Experiment*

We ran the online experiment using MTurk. We first asked MTurk subjects to transcribe three paragraphs of text to reduce the likelihood of their dropping from the study after seeing treatment. After the lock-in task, subjects have an opportunity to split a 50 cent bonus (separate from the payment they received for data entry) with the charitable recipient, the Red Cross. We believed the Red Cross to be more well-known for MTurk subjects, who come mostly from the U.S. and India. Workers then provided their gender, age, country of residence, religion, and how often they attend

religious services. We had 902 decisions from 902 subjects.[15]

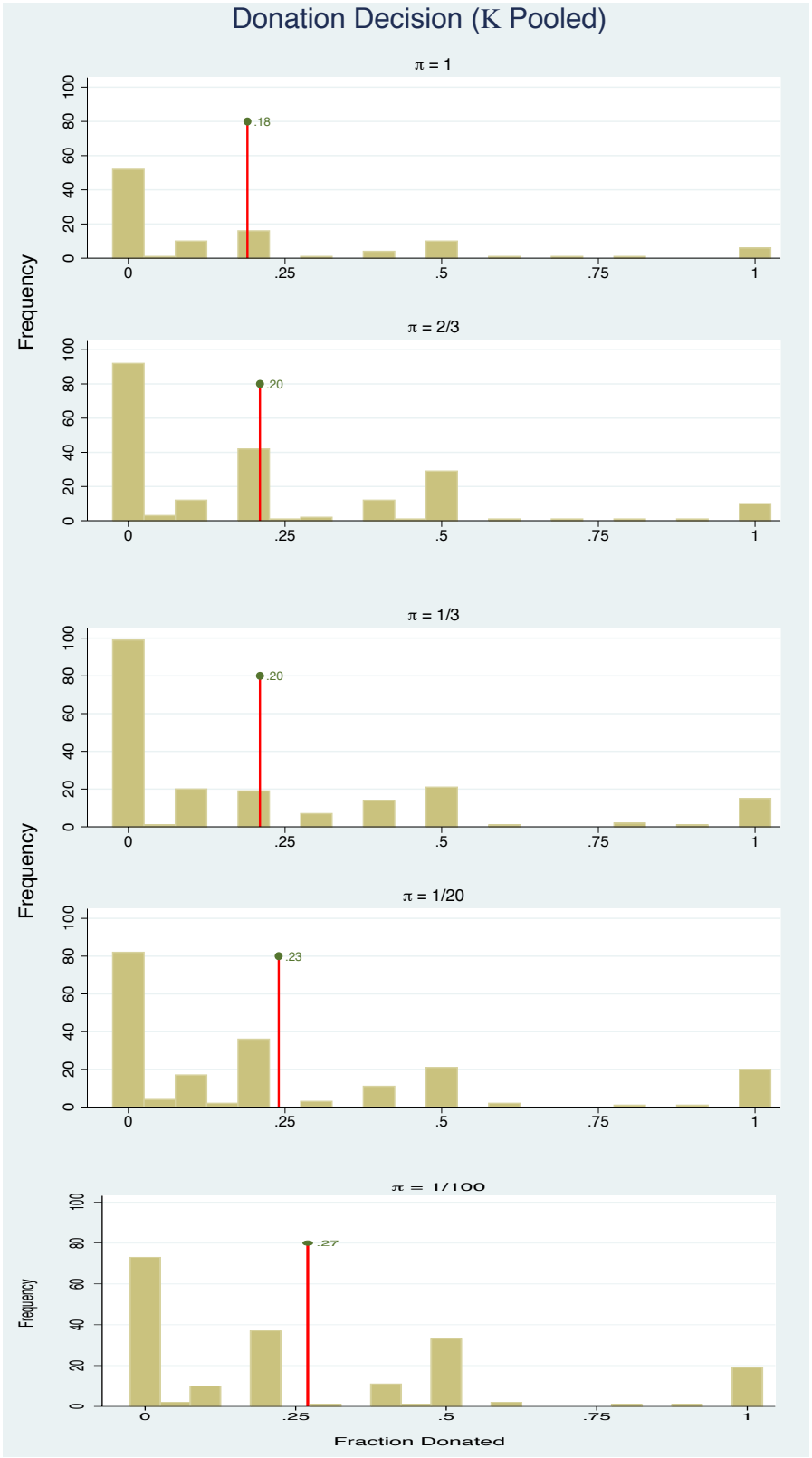
Participants were randomly assigned to one of five groups with π being: 100%, 66%, 33%, 5%, and 1%. They were told in advance about the implementation probability. We randomized such that we collected roughly 200 subjects in each of the 66%, 33%, 5%, and 1% treatments and 100 subjects in the 100% treatment. In addition, we randomize κ to be 50 cents (maximum) and 0 cents (minimum). Appendix C presents instructions.[16] All our analyses are reported in terms of fraction donated from 0 to 1.

Figure 9 shows that the lower the π, the more generous is the decision-maker. The increase in generosity is monotonic with the decrease in probability. Donations increased from 18% (when $\pi = 1$) to 27% (when $\pi = 0.01$). The following presents regression results and we can again strongly reject the hypothesis that subjects are targeting expected income or expected giving.

[15] 2 individuals did not report a complete set of demographic characteristics, so they are dropped in some of the regressions.

[16] To assess potential anchoring effects induced by κ, we also ran an auxiliary experiment that randomized κ to be 10 cents or unknown to workers (they are told the computer is making a determination) and we draw κ from a uniform distribution between 0 and 50. When κ was unknown, we also asked workers what they believed would be the amount donated if the computer made the decision. We found that 18% of subjects gave 10 cents in the "$\kappa = 10$ Cents" treatment while 14% gave 10 cents in the "$\kappa =$ Unknown" treatment. Since we did not see significant anchoring effects, it is not the focus of our analysis.

FIGURE 9.— Donation and π: Raw data (MTurk)



Red: Mean

Table III reports that the effect of π is significant at the 5% level in a linear regression in Column 1. The effect size of 7.2% is roughly one-third of the mean donation of 23%. Column 2 adds demographic controls.[17] The point estimates are stable. Columns 3 and 4 consider if subjects target expected income and Columns 5 and 6 consider expected giving. We can strongly reject the hypothesis that subjects are targeting these quantities. Increasing the likelihood of implementation from 0 to 1 reduces the expected income of the donee by 22% and increases the expected giving of the donor by 20%.[18]

TABLE III

DONATION AND π: LINEAR REGRESSION (MTURK)

	Ordinary Least Squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected Income E(x_2)		Expected Giving (πd^*)	
Mean dep. var.	0.23		0.34		0.07	
% Consequential (π)	-0.0725**	-0.0684*	-0.224***	-0.219***	0.194***	0.213***
	(0.0288)	(0.0390)	(0.0334)	(0.0299)	(0.0132)	(0.0181)
K Fixed Effects	N	Y	N	Y	N	Y
Controls	N	Y	N	Y	N	Y
Observations	902	900	902	900	902	900
R-squared	0.007	0.059	0.048	0.604	0.194	0.214

Notes: Standard errors in parentheses. Raw data shown in Figure 10. Controls include indicator variables for gender, American, Indian, Christian, Atheist, aged 25 or younger, and aged 26-35 as well as continuous measures for religious attendance and accuracy in the lock-in data entry task. * p < 0.10, ** p < 0.05, *** p < 0.01

Table IV presents separate linear regressions for each κ treatment-arm. In each pair of columns (without controls and with controls), we find a quantitatively similar 5.3% to 7.8% decrease as π goes from 0 to 1. The effects are not significantly different across treatment arms.

[17] Country of origin was coded as United States and India with the omitted category as other; religion was coded as Christian, Hindu, and Atheist with the omitted category as other; religious services attendance was coded as never, once a year, once a month, once a week, or multiple times a week.

[18] To make calculations on expected donations when κ is unknown, we use data on perceived donation.

TABLE IV

DONATION AND π: LINEAR REGRESSION DISAGGREGATED BY κ (MTURK)

	Ordinary Least Squares							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Decision (d)		Decision (d)		Decision (d)		Decision (d)	
	K = Unknown		K = 10¢		K = 0¢		K = 50¢	
Mean dep. var.	0.26		0.22		0.20		0.22	
% Consequential (π)	-0.0778	-0.0654	-0.0525	-0.0321	-0.0711	-0.0708	-0.0644	-0.0675
	(0.0523)	(0.0523)	(0.0526)	(0.0536)	(0.0464)	(0.0466)	(0.0462)	(0.0456)
Male		-0.0909**		-0.0474		0.0108		0.0178
		(0.0399)		(0.0430)		(0.0395)		(0.0362)
American		0.0241		-0.0539		0.0838		0.117*
		(0.0524)		(0.0539)		(0.0664)		(0.0598)
Indian		-0.0672		-0.0785		-0.0673		-0.0626
		(0.0566)		(0.0560)		(0.0630)		(0.0590)
Christian		-0.0295		0.0584		-0.0215		-0.000293
		(0.0483)		(0.0503)		(0.0494)		(0.0479)
Atheist		-0.0188		0.00480		0.0113		-0.0927
		(0.0644)		(0.0649)		(0.0802)		(0.0725)
Religious Services Attendance		-0.00614		0.000508		0.00367		-0.00546
		(0.0145)		(0.0156)		(0.0137)		(0.0137)
Ages 25 or Under		-0.0207		-0.122**		-0.0109		-0.113**
		(0.0518)		(0.0570)		(0.0493)		(0.0474)
Ages 26-35		0.00271		-0.110*		-0.00105		-0.111**
		(0.0548)		(0.0593)		(0.0493)		(0.0480)
Own Errors		-0.000192		-0.000186		0.000220		-0.000148
		(0.000193)		(0.000163)		(0.000194)		(0.000143)
Observations	260	260	218	218	256	255	271	270
R-squared	0.009	0.069	0.005	0.081	0.009	0.052	0.007	0.097

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

We next examine whether the distributions of donation decisions are significantly affected by π. Table V shows that along most thresholds for π, Mann-Whitney tests yield significant differences in the distribution of donations as π increases. To interpret, 0.05 in Column 1 means that we reject with 95% confidence the hypothesis that the distribution of decisions for subjects treated with $\pi = 1, 0.67, 0.33$ is the same as the distribution of decisions for subjects treated with $\pi = 0.05, 0.01$. The lower panel of Table V reports that the distribution of donations does not significantly vary by κ. Means are also not significantly different by κ.

TABLE V

DONATION AND π: NON-PARAMETRIC TESTS (MTURK)

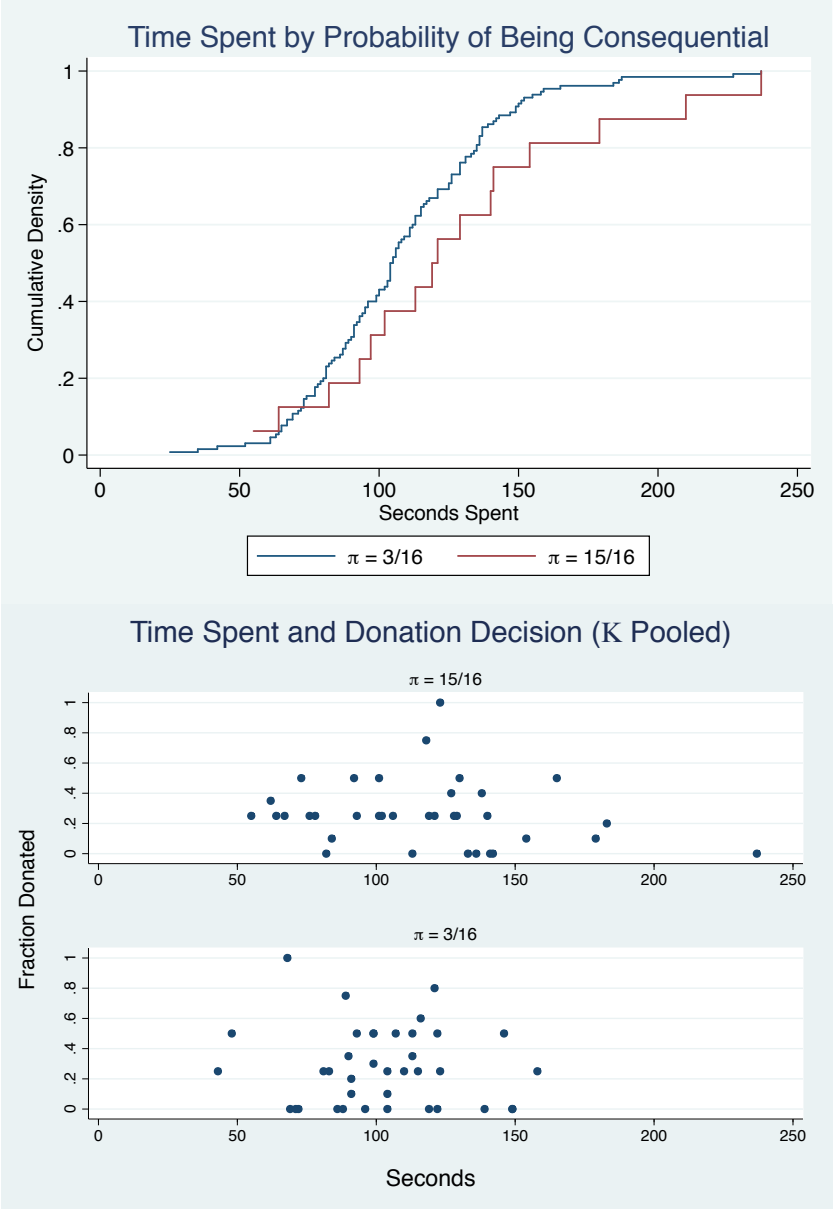
	Wilcoxon-Mann-Whitney 2-sided test (p-values)		
	(1)	(2)	(3)
Thresholds	K Unknown or 10¢	$K = 0$¢ or 50¢	K Pooled
$\pi = 1$ vs. $\pi \leq 0.67$	0.91	0.05	0.11
$\pi \geq 0.67$ vs. $\pi \leq 0.33$	0.07	1.00	0.20
$\pi \geq 0.33$ vs. $\pi \leq 0.05$	0.05	0.10	0.01
$\pi \geq 0.05$ vs. $\pi = 0.01$	0.15	0.02	0.01

	π Pooled
$K \geq 10$¢ vs. $K = 0$¢	0.40
$K = 50$¢ vs. $K \leq 10$¢	0.11

Next, we reject cognition costs as the driving feature for decision change. The three findings are 1) individuals spend roughly the same time thinking about their decision regardless of the implementation probability, 2) donations were not associated with time spent, and 3) those estimated to be most responsive to implementation probability do not seem to be resorting to heuristics more, at least measured by time spent.

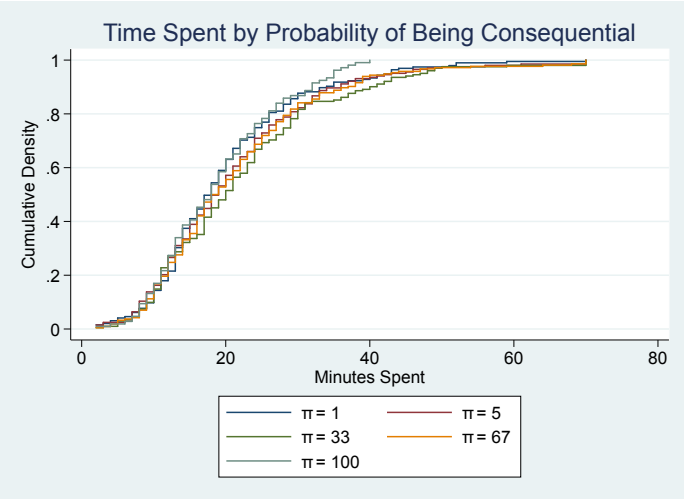
Figure 10 shows that individuals spend roughly the same time thinking about their decision regardless of the implementation probability, which is inconsistent with the cognitive cost model, where individuals spend less time thinking and use altruistic heuristics when their decision is less likely to be implemented. Moreover, subjects do not donate less when they spend more time on their decision to compensate for cognition effort.

FIGURE 10.— Time Spent (on Donation Decision): Lab



On MTurk, we did not have data on the time spent before and after the donation decision and only had data for the entire MTurk session, which is displayed in Figure 11. We find that time spent is only affected (and *reduced*) by $\pi = 1$. This result would appear inconsistent with a cognition costs theory where individuals spend more time on decisions when they are consequential. Donations were again not associated with time spent, but would be negatively associated under

FIGURE 11.— Time Spent (Begin vs. End Time): MTurk



a theory that cognition costs explain increased generosity when the implementation probability is low.

Table VI shows that at low π, those with below-median $\frac{\delta d}{\delta \pi}$ spend less time than those with above-median $\frac{\delta d}{\delta \pi}$ (see below for an explanation for how these groups are determined). In addition, Figure 12 shows that those with high $\frac{\delta d}{\delta \pi}$ do not vary their time spent as π changes. These findings are inconsistent with the cognition cost model in that those whose behaviors are most elastic to π (high $\frac{\delta d}{\delta \pi}$) do not seem to be resorting to heuristics more when the probability of being consequential is low, at least measured by time spent.

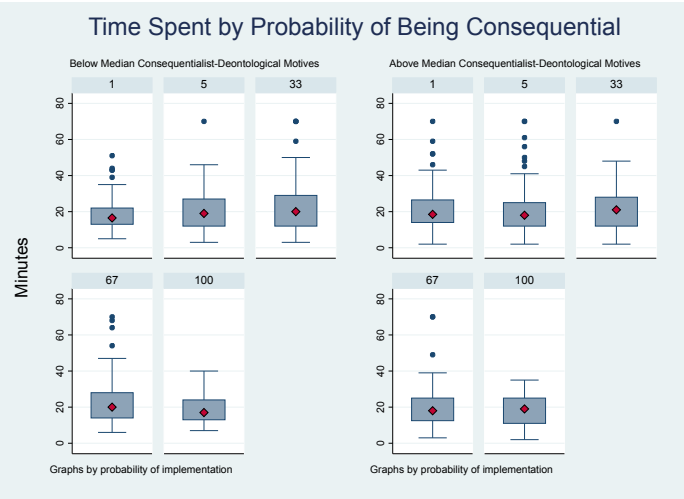
TABLE VI

TIME SPENT (BEGIN VS. END TIME): MTURK HETEROGENEITY BY $\frac{\delta d}{\delta \pi}$

Sample	All Subjects	Above Median Mixed-Consequentialist		Below Median Mixed-Consequentialist	
	(1)	(2)	(3)*	(4)	(5)*
Mean dep. var.			20.8		
% Consequential (π)	0.0123	0.0176	0.0452	0.163***	0.118*
	(0.0162)	(0.0547)	(0.0574)	(0.0548)	(0.0635)
π^2		-0.000482	-0.000452	-0.00167***	-0.00122*
		(0.000573)	(0.000602)	(0.000581)	(0.000674)
Above Median Mixed-Consequentialist	0.755				
	(1.119)				
π * Above Median Mixed-Consequentialist	-0.0386*				
	(0.0227)				
Observations	900	449	449	451	451
R-squared	0.004	0.008		0.019	

Notes: Standard errors in parentheses. Mixed-Consequentialist aggregates for each subject their demographic characteristics' contribution to the effect of π on the Donation decision. Regressions are weighted by the standard deviation of the first regression to account for uncertainty in the calculation of mixed-consequentialist score. Columns 3 and 5 employ median regressions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE 12.— Time Spent by $\frac{\delta d}{\delta \pi}$: MTurk



Red Diamond: Median

To estimate high and low $\frac{\delta d}{\delta \pi}$ and to explore sensitivity of the decision d to π, we construct synthetic cohorts. Formally, we estimate:

$$Donation_i = \beta_0 \pi_i + \beta_1 \mathbf{X}_i \pi_i + \alpha \mathbf{X}_i + \varepsilon_i$$

We interpret the change in d to π as measuring the mixed consequentialist-deontological motives. We then compute for each individual:

$$MixedConsequentialistDeontological_i = |\hat{\beta}_0 + \widehat{\beta}_1 \mathbf{X}_i|$$

We use all the demographic characteristics in $\mathbf{X}_i$ to construct the mixed consequentialist-deontological score. Each subject's demographic characteristics are then used to calculate a predicted mixed consequentialist-deontological score by taking the absolute value of the sum of the contributions of their demographic characteristics along with the constant term.

Table VII shows that along *all* demographic groups, $\frac{\delta d}{\delta \pi} < 0$. Americans, Christians, Atheists, and those who are less likely to attend religious services are particularly likely to have steeper $\frac{\delta d}{\delta \pi}$.

TABLE VII

WHO RESPONDS TO π? (AMT)

	Ordinary Least Squares									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Decision (d)									
Mean dep. var.	0.23									
% Consequential (π)	-0.100**	-0.0493	-0.124**	-0.0500	-0.0522	-0.0774	-0.0618	-0.0548	-0.0839**	-0.0190
	(0.0494)	(0.0429)	(0.0506)	(0.0436)	(0.0403)	(0.0616)	(0.0467)	(0.0443)	(0.0407)	(0.126)
π * Male	0.0612									0.0490
	(0.0577)									(0.0611)
π * American		-0.0675								0.0370
		(0.0627)								(0.0911)
π * Indian			0.0990*							0.0426
			(0.0574)							(0.0963)
π * Christian				-0.0599						-0.0658
				(0.0632)						(0.0783)
π * Atheist					-0.133					-0.145
					(0.0837)					(0.108)
π * Religious Services Attendance						0.00394				-0.00739
						(0.0210)				(0.0224)
π * Ages 25 or Under							-0.0149			-0.0815
							(0.0576)			(0.0787)
π * Ages 26-35								-0.0386		-0.0878
								(0.0597)		(0.0808)
π * Own Errors									0.000402	0.000319
									(0.000299)	(0.000307)
K Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	900	900	900	900	900	900	900	900	900	900
R-squared	0.061	0.061	0.063	0.060	0.062	0.059	0.059	0.060	0.061	0.068

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

4.3. *Structural Estimation*

This section presents structural estimates of how individuals trade off between consequentialist and deontological motivations. We provide two illustrations. First, we follow Cappelen et al. (2007) and Cappelen et al. (2013) and assume that homogenous individuals maximize homo oeconomicus consequentialist motivations, but place weight λ on a deontological portion that follows bliss point preferences: $u(x_{DM,}, x_2, d) = \lambda(x_1) + (-(\delta - d)^2) = \lambda(1-d) + (-(\delta - d)^2)$.[19] The first-order condition is $0 = \pi\lambda(-1) + 2(\delta - d)$, which results in a linear regression, $-\frac{\lambda}{2}\pi + \delta = d^*$.

Note that we can interpret the constant term of the linear regression as the bliss point, representing the decision when $\pi = 0$. Figure 9 would yield a bliss point $\delta = 0.25$, which is very close to the observed 27% when $\pi = 0.01$. Then, since we can pin down one of two unknown parameters, we can identify the weight placed on deontological motivations using the speed of change as π varies; in this case, $\lambda = 0.14$. Note that a pure homo oeconomicus would maximize d^* at 0, which is why λ increases monotonically with speed of change.

[19]Note this means that a Cappellen et al. model views duty as $d = \delta$ rather than $d \geq \delta$. We assume that subjects' duties are enumerated in percent terms.

TABLE VIII

DONATION AND π: LINEAR REGRESSION

	OLS	IV	IV
	(1)	(2)	(3)
	Decision (d)		
Mean dep. var.	0.23		
% Consequential (π)	-0.239***	-0.363***	-0.368***
	(0.0249)	(0.0548)	(0.139)
π * $1(d \geq w/2)$	0.870***	1.516***	1.542**
	(0.0412)	(0.250)	(0.714)
Constant (Duty Bliss Point)	0.251***	0.249***	0.249***
	(0.0116)	(0.0131)	(0.0134)
IV	N	π, Indian	π, Age $\leq$ 25
Observations	902	902	902
R-squared	0.336	0.155	0.140

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Our second illustration models consequentialist motivations as Fehr and Schmidt (1999), plugging in α and β inequality parameters for $u(x_{DM,}, x_2, d) = \lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (-(\delta - d)^2)$. The individual's first-order condition over their choice d is then given by the following expression: If $\frac{1}{2} > d$, then $0 = \pi\lambda(2\beta - 1) + 2(\delta - d)$, else $0 = \pi\lambda(-2\alpha - 1) + 2(\delta - d)$.[20] Thus, we run a linear regression of d on $1[\frac{1}{2} > d]\pi$ and $1[\frac{1}{2} \leq d]\pi$. We present estimates using two different instruments for $1[\frac{1}{2} \leq d_i]$, which results in similar point estimates (Table VIII).[21]

The bliss point is still 25%. Then, the first coefficient in the regression model indicates that while $d < 50\%$, donation increases as π decreases. However, once $d > 50\%$, donation decreases as π decreases. This switch is intuitive because the bliss point for duty is below 50% and we still assume Cappelen et al. bliss point preferences. As π falls, they should move towards the bliss point, which is less than 50%. Our coefficients also have a structural interpretation for λ. Table VIII yields $\frac{\lambda(2\beta-1)}{2} = -0.36$ and $\frac{\lambda(-2\alpha-1)}{2} = 1.16$. Finally, we need to make an assumption for α and β. For

[20] We model consequentialist motivations as Fehr and Schmidt (1999), plugging in α and β inequality parameters for $u(x_{DM,}, x_2, d) = \lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (-(\delta - d)^2)$. The individual's first-order condition over their choice d is then given by the following expression: If $\frac{1}{2} > d$, then $0 = \pi\lambda(2\beta - 1) + 2(\delta - d)$, else $0 = \pi\lambda(-2\alpha - 1) + 2(\delta - d)$.

The derivation is as follows: $\pi\lambda(1 - d - \alpha max\{2d - 1, 0\} - \beta max\{1 - 2d, 0\}) + (-(\delta - d)^2)$. This expression is quadratic in d, so the first-order condition, and hence moment conditions, will be linear in d. Thus, we estimate a linear regression to back out our parameters of interest. To see this, first observe that the decision-dependent portion of expected utility if $\frac{1}{2} > d$, is: $\pi\lambda(1 - d - \beta(1 - 2d)) + (-(\delta - d)^2)$, else $\pi\lambda(1 - d - \alpha(2d - 1)) + (-(\delta - d)^2)$. Thus, our linear regression is: If $\frac{1}{2} > d$, then $\pi\frac{\lambda(2\beta-1)}{2} + \delta = d^*$, else $\pi\frac{\lambda(-2\alpha-1)}{2} + \delta = d^*$. This expression motivates our GMM condition:

$E\left[\pi\left(1[\frac{1}{2} > d]\left[d - \pi\frac{\lambda(2\beta-1)}{2} - \delta\right] + 1[\frac{1}{2} \leq d]\left[d - \pi\frac{\lambda(-2\alpha-1)}{2} - \delta\right]\right)\right] = 0.$

[21] The OLS regression is also presented, but problematic because the decision appears on the left-side of the equation as outcome and on the right-side in the indicator function. The data limits our choice of instruments for $1[\frac{1}{2} \leq d_i]$ that is not directly correlated with d_i. The instruments are being under 25 or being from India.

the range of plausible α and β values in Fehr and Schmidt (1999), our data is inconsistent with the joint hypothesis of consequentialist motivations being Fehr-Schmidt, the duty motivation being bliss point, and a non-zero weight on consequentialist motivations.[22] Taken together, each of the three exercises offer unique advantages and limitations that portray a picture of variance in response to the probability of implementation.

5. CONCLUSION

Recent advances in economic theory, motivated by experimental findings, have led to the adoption of models where individuals make decisions not solely based on self-interest (considering consequences for oneself), but also based on the consequences for others. Investigations of motives over decisions *per se*, independently of their consequences, are rare. In this paper, we formalize the notion of consequentialist as well as deontological motivations as properties of preference relations; we suggest and implement a thought experiment that uses revealed preference to detect deontological motivations—varying the probability that one's decision is consequential (i.e., implemented). For a consequentialist who satisfies first order stochastic dominance, the optimal decision is independent of the probability that the action will be enacted. For a deontologist, the optimal decision is also independent of the probability. Only mixtures of both consequentialist and deontological motivations predict changes in behavior as the probability changes.

Our research design has some implications for the random lottery method in experimental economics. Prior formal observations support its use—roughly speaking, if individuals satisfy the independence axiom (Holt 1986), then the random lottery method is valid—and these theoretical observations have been empirically validated (Starmer and Sugden 1991; Hey and Lee 2005). What we show is that when it comes to decisions that are not purely economic (e.g., social preference decisions that can have a deontological motive), if individuals satisfy FOSD, the random lottery method can reveal different decisions that are more pro-social than when the decisions are consequential.

Future research may explore several legal applications. First, measuring *intent* in law, most famously, in criminal law when a distinction is made between *mens rea* (intention) and *actus reus* (act): did the shooter *intend* to kill (but didn't) *or* did the shooter unintentionally commit the *act*

[22] With two equations and three unknowns, we cannot identify our parameters. However, we can choose values for β and α in the range of values in Fehr and Schmidt (1999). But, if individuals are inequality averse and are more averse to adverse inequality, we know that $\alpha > \beta > 0$; examining $\frac{\lambda(-2\alpha-1)}{2} = 1.16$ implies $\lambda < 0$, but $\lambda = 0$ is a boundary condition.

of killing. In other instances, the law also cares about mental states beyond just the consequences, such as the litigant's motivations in copyright disputes, where a litigant has cause of action only if she is motivated by her *moral rights* to litigate, that is, she is not litigating because of the consequences of winning. More broadly, in equity law, judges may care about opportunistic behavior as opposed to the behavior itself, which is similar to the decision-maker having both *mens rea* and *actus reus.* Finally, some philosophers argue that human dignity derives from the possibility of deontological decision-making—"what commands respect is the capacity for morality" (Waldron 2012) and "Everything has either a price or a dignity. What has a price can be replaced by something else as its equivalent; what, on the other hand, is raised above all price and therefore admits of no equivalent has a dignity. .. humanity insofar as it is capable of morality is that which alone has dignity" (Kant 1797).

REFERENCES

Akerlof, George A., and Rachel E. Kranton, 2000, Economics and Identity, *The Quarterly Journal of Economics* 115, 715–753.

Alexander, Larry, and Michael Moore, 2012, *Stanford Encyclopedia of Philosophy*.

Alger, Ingela, and Jörgen W Weibull, 2013, Homo Moralis-Preference Evolution Under Incomplete Information and Assortative Matching, *Econometrica* 81, 2269–2302.

Andreoni, James, 1990, Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving, *The Economic Journal* 100, 464–477.

Andreoni, James, and B. Douglas Bernheim, 2009, Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects, *Econometrica* 77, 1607–1636.

Anscombe, Francis J, and Robert J Aumann, 1963, A definition of subjective probability, *Annals of mathematical statistics* 199–205.

Arrow, Kenneth Joseph, 2012, *Social Choice and Individual Values*, Cowles Foundation Monographs Series, third edition (Yale university press, New Haven), Monograph 12.

Batson, C. Daniel, Judy G. Batson, Jacqueline K. Slingsby, Kevin L. Harrell, Heli M. Peekna, and R. Matthew Todd, 1991, Empathic Joy and the Empathy-Altruism Hypothesis, *Journal of Personality and Social Psychology* 61, 413–426.

Battigalli, Pierpaolo, and Martin Dufwenberg, 2007, Guilt in Games, *The American Economic Review* 97, 170–176.

Bénabou, Roland, and Jean Tirole, 2006, Incentives and Prosocial Behavior, *The American Economic Review* 96, 1652–1678.

Bénabou, Roland, and Jean Tirole, 2011, Identity, Morals, and Taboos: Beliefs as Assets, *The Quarterly Journal of Economics* 126, 805–855.

Bentham, Jeremy, 1791, *Panopticon* (T. Payne, London).

Bergstrom, Theodore C., Rodney J. Garratt, and Damien Sheehan-Connor, 2009, One Chance in a Million: Altruism and the Bone Marrow Registry, *The American Economic Review* 99, 1309–1334.

Besley, Timothy, 2005, Political selection, *The Journal of Economic Perspectives* 19, 43–60.

Bigenho, Jason, and Seung-Keun Martinez, 2019, Social Comparisons in Peer Effects, Technical report, UCSD.

Binmore, Ken, 1994, Playing fair: Game theory and the social contract, *Cambridge, Mass.: MIT Press* .

Bowles, Samuel, and Sandra Polania-Reyes, 2012, Economic Incentives and Social Preferences: Substitutes or Complements?, *Journal of Economic Literature* 50, 368–425.

Brock, J. Michelle, Andreas Lange, and Erkut Y. Ozbay, 2013, Dictating the Risk: Experimental Evidence on Giving in Risky Environments, *The American Economic Review* 103, 415–437.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø Sørensen, and Bertil Tungodden, 2007, The Pluralism of Fairness Ideals: An Experimental Approach, *American Economic Review* 97, 818–827.

Cappelen, Alexander W, James Konow, Erik Ø Sørensen, and Bertil Tungodden, 2013, Just luck: An experimental study of risk-taking and fairness, *The American Economic Review* 103, 1398–1413.

Cavaille, Charlotte, Daniel L. Chen, and Karin Van der Straeten, 2018, Who Cares? Measuring Attitude Strength in a Polarized Environment .

Chen, Daniel, and J.J. Prescott, 2016, Implicit Egoism in Sentencing Decisions: First Letter Name Effects with Randomly Assigned Defendants, Technical report.

Chen, Daniel L., Moti Michaeli, and Daniel Spiro, 2015, Ideological Perfectionism on Judicial Panels, Working paper, ETH Zurich.

Chen, Daniel L., and Martin Schonger, 2015, A Theory of Experiments: Invariance of Equilibrium to the Strategy Method of Elicitation and Implications for Social Preferences, Working paper, ETH Zurich.

Chlaß, Nadine, Werner Güth, Topi Miettinen, et al., 2014, Purely procedural preferences-beyond procedural equity and reciprocity, Technical report, Stockholm School of Economics, Stockholm Institute of Transition Economics.

Choi, Hyunkyung, Marcia Van Riper, and Suzanne Thoyre, 2012, Decision making following a prenatal diagnosis of Down syndrome: an integrative review, *Journal of Midwifery & Womens Health* 57, 156–164.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman, 2014, Who is (More) Rational?, *The American Economic Review* 104, 1518–1550.

Cilliers, Jacobus, Oeindrila Dube, and Bilal Siddiqi, 2015, The White-Man Effect: How Foreigner Presence Affects Behavior in Experiments, *Journal of Economic Behavior and Organization* .

Dana, Jason, Daylian M Cain, Robyn M Dawes, et al., 2006, What you don't know won't hurt me: Costly (but quiet) exit in dictator games, *Organizational Behavior and Human Decision Processes* 100, 193–201.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, 2007, Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness, *Economic Theory* 33, 67–80, Symposium on Behavioral Game Theory.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier, 2013, Voting to Tell Others, Working paper.

Elizabeth Hoffman, Matthew L. Spitzer, 1985, Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice, *The Journal of Legal Studies* 14, 259–297.

Ellingsen, Tore, and Magnus Johannesson, 2008, Pride and Prejudice: The Human Side of Incentive Theory, *The American Economic Review* 98, 990–1008.

Falk, Armin, and Urs Fischbacher, 2006, A Theory of Reciprocity, *Games and Economic Behavior* 54, 293–315.

Falk, Armin, and Nora Szech, 2013, Morals and Markets, *Science* 340, 707–711.

Feddersen, Timothy, Sean Gailmard, and Alvaro Sandroni, 2009, Moral Bias in Large Elections: Theory and Experimental Evidence, *The American Political Science Review* 103, 175–192.

Fehr, Ernst, and Klaus M. Schmidt, 1999, A Theory of Fairness, Competition, and Cooperation, *The Quarterly Journal of Economics* 114, 817–868.

Fischbacher, Urs, 2007, z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental economics* 10, 171–178.

Foot, Philippa, 1967, The Problem of Abortion and the Doctrine of Double Effect, *Oxford Review* 5, 5–15.

Friedman, Milton, and Leonard J. Savage, 1948, The Utility Analysis of Choices Involving Risk, *The Journal of Political Economy* 56, 279–304.

Gibson, Rajna, Carmen Tanner, and Alexander F. Wagner, 2013, Preferences for Truthfulness: Heterogeneity among and within Individuals, *The American Economic Review* 103, 532–548.

Gneezy, Uri, 2005, Deception: The Role of Consequences, *The American Economic Review* 95, 384–394.

Grossman, Zachary, 2015, Self-signaling and social-signaling in giving, *Journal of Economic Behavior & Organization* 117, 26–39.

Harsanyi, John C, 1977, Morality and the theory of rational behavior, *Social Research* 623–656.

Hey, John D, and Jinkwon Lee, 2005, Do subjects separate (or are they sophisticated)?, *Experimental Economics* 8, 233–265.

Holt, Charles A., 1986, Preference Reversals and the Independence Axiom, *The American Economic Review* 76, 508–515.

Junker-Kenny, Maureen, 2013, Recognising traditions of argumentation in philosophical ethics, in *Ethics for Graduate Researchers*, 7–26 (Elsevier).

Kant, Immanuel, 1797, Über ein vermeintes Recht aus Menschenliebe zu lügen, *Berlinische Blätter* 1, 301–314.

Kaplow, Louis, and Steven Shavell, 2006, *Fairness Versus Welfare* (Harvard University Press).

Konow, James, 2000, Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions, *The American Economic Review* 90, 1072–1091.

Krawczyk, Michal Wiktor, 2011, A model of procedural and distributive fairness, *Theory and Decision* 70, 111–128.

Kreps, David M, 1988, *Notes on the Theory of Choice* (Westview Press Boulder).

Levhari, David, Jacob Paroush, and Bezalel Peleg, 1975, Efficiency Analysis for Multivariate Distributions, *The Review of Economic Studies* 42, 87–91.

Machina, Mark J., 1982, "Expected Utility" Analysis without the Independence Axiom, *Econometrica* 50, 277–323.

Machina, Mark J., 1989, Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty, *Journal of Economic Literature* 27, 1622–1668.

Mankiw, N. Gregory, and Matthew Weinzierl, 2010, The Optimal Taxation of Height: A Case Study of Utilitarian Income Redistribution, *American Economic Journal: Economic Policy* 2, 155–176.

McCabe, Kevin A., Mary L. Rigdon, and Vernon L. Smith, 2003, Positive reciprocity and intentions in trust games, *Journal of Economic Behavior & Organization* 52, 267–275.

Nozick, Robert, 1974, *Anarchy, State, and Utopia*, Harper Torchbooks (Basic Books).

Quiggin, John, 1982, A Theory of Anticipated Utility, *Journal of Economic Behavior & Organization* 3, 323–343.

Quiggin, John, 1990, Stochastic Dominance in Regret Theory, *The Review of Economic Studies* 57, 503–511.

Rabin, Matthew, 1993, Incorporating Fairness into Game Theory and Economics, *The American Economic Review* 83, 1281–1302.

Rand, David G., Joshua D. Greene, and Martin A. Nowak, 2012, Spontaneous Giving and Calculated Greed, *Nature* 489, 427–430.

Riker, William H., and Peter C. Ordeshook, 1968, A Theory of the Calculus of Voting, *The American Political Science Review* 62, 25–42.

Roth, Alvin E., 2007, Repugnance as a Constraint on Markets, *The Journal of Economic Perspectives* 21, 37–58.

Savage, Leonard J, 1972, *The Foundations of Statistics* (Courier Corporation).

Shayo, Moses, and Alon Harel, 2012, Non-consequentialist voting, *Journal of Economic Behavior & Organization* 81, 299–313.

Sinnott-Armstrong, Walter, 2012, Consequentialism, in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*.

Smith, Adam, 1761, *The Theory of Moral Sentiments* (A. Millar).

Smith, Kyle D., John P. Keating, and Ezra Stotland, 1989, Altruism Reconsidered: The Effect of Denying Feedback on a Victim's Status to Empathic Witnesses, *Journal of Personality and Social Psychology* 57, 641–650.

Sobel, Joel, 2005, Interdependent preferences and reciprocity, *Journal of economic literature* 43, 392–436.

Starmer, Chris, 2000, Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk, *Journal of Economic Literature* 38, 332–382.

Starmer, Chris, and Robert Sugden, 1991, Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation, *The American Economic Review* 81, 971–978.

Tetlock, Philip E, 2003, Thinking the unthinkable: Sacred values and taboo cognitions, *Trends in cognitive sciences* 7, 320–324.

Trautmann, Stefan T, 2009, A tractable model of process fairness under risk, *Journal of Economic Psychology* 30, 803–813.

Tversky, Amos, and Daniel Kahneman, 1992, Advances in Prospect Theory: Cumulative Representation of Uncertainty, *Journal of Risk and Uncertainty* 5, 297–323.

Tyler, Tom R, 1997, The psychology of legitimacy: A relational perspective on voluntary deference to authorities, *Personality and social psychology review* 1, 323–345.

Wakker, Peter, 1993, Savage's Axioms Usually Imply Violation of Strict Stochastic Dominance, *The Review of Economic Studies* 60, 487–493.

Waldron, Jeremy, 2012, *Dignity, rank, and rights* (Oxford University Press).

Wilcox, Nathaniel T., 1993, Lottery Choice: Incentives, Complexity and Decision Time, *The Economic Journal* 103, 1397–1417.

For Online Publication

Web Appendix:

APPENDIX A: FORMAL STATEMENT OF ASSUMPTIONS AND THEOREM

This appendix provides additional details and proofs for our results. For completeness, we also include all relevant information from the main text in this appendix.

The standard consequentialist approach to choice under uncertainty (and central assumption for choice behavior regarding uncertainty) is first-order stochastic dominance (FOSD). A wide variety of models of choice under uncertainty satisfies FOSD and thus falls within this framework, among them most prominently, expected utility theory, its generalization by Machina (1982), but also cumulative prospect theory (Tversky and Kahneman 1992) or rank-dependent utility theory (Quiggin 1982). Stochastic dominance is a compelling criterion for decision-making quality and is generally accepted in decision theory (Quiggin 1990; Wakker 1993; Choi et al. 2014).

In the following paragraph and the axioms up to FOSD, we closely follow the canonical framework as laid out in Kreps (1988). Let there be outcomes x. x can be a real valued vector. In the thought experiment, it would be $x = (x_1, x_2)$. Let the set of all x be finite and denote it by X. A probability measure on X is a function $p : X \to [0, 1]$ such that $\sum_{x \in X} p(x) = 1$. Let P be the set of all probability measures on X, and therefore, in the thought experiment, a subset of it, is the choice set of the decision-maker.

Axiom 2 *(preference relation) Let $\succsim$ be a complete and transitive preference on P.*

Axiom 1 is the standard one saying that the preference relation is a complete ordering. It implicitly includes consequentialism since the preference relation is on P, that is, over lotteries that are over consequences x.

Next we define first-order stochastic dominance (FOSD). Often, definitions of FOSD are suitable only for preference relations that are monotonic in the real numbers, for example see Levhari et al. (1975). These definitions define FOSD with respect to the ordering induced by the real numbers, assuming that prices are vectors. It is important to define FOSD with respect to ordering over outcomes rather than the outcomes themselves.[23]

Definition (FOSD) p first-order stochastically dominates q with respect to the ordering induced by $\succsim$, if for all x':

$\sum_{x:x' \succsim x} p(x) \leq \sum_{x:x' \succsim x} q(x)$.

Axiom *(FOSD) If p FOSD q with respect to the ordering induced by $\succsim$, then $p \succsim q$.*

Formally, our theorem needs both strict FOSD and weak FOSD since the former does not imply the latter.

Definition (Strict FOSD) p strictly first-order stochastically dominates q with respect to the ordering induced by $\succsim$ if p FOSD q with respect to that ordering, and there exists an x' such that:

$\sum_{x:x' \succsim x} p(x) < \sum_{x:x' \succsim x} q(x)$.

Axiom *(Strict FOSD) If p strictly FOSD q with respect to the ordering induced by $\succsim$, then $p \succ q$.*

[23] FOSD over outcomes is inappropriate in the context of social preferences, which are often not monotonic due to envy or fairness concerns.

The following theorem implies that in our thought experiment, changing the probability of being consequential π does not change the decision. It is this prediction of the theory that we test and interpret a rejection of the prediction as evidence that people are not purely consequentialist.

THEOREM 2 *If the DM satisfies the axioms Preference Relation, FOSD, and Strict FOSD, and there exist* $x, x', x'' \in X'$ *and* $\pi\epsilon(0;1]$ *such that* $\pi x + (1-\pi)x'' \succcurlyeq \pi x' + (1-\pi)x''$*, then for all* $\pi'\epsilon(0;1] : \pi' x + (1-\pi')x'' \succcurlyeq \pi' x' + (1-\pi')x''$.

PROOF: (i) $x \succsim x'$: Suppose not, then $x' \succ x$, and therefore $\pi x' + (1-\pi)x''$ strongly first-order stochastically dominates $\pi x + (1-\pi)x''$. Then by axiom Strict FOSD, $\pi x' + (1-\pi)x'' \succ \pi x + (1-\pi)x''$, a contradiction.

(ii) Since $x \succsim x'$, $\pi' x + (1-\pi')x''$, first-order stochastically dominates $\pi' x' + (1-\pi')x''$. Thus by axiom FOSD, $\pi' x + (1-\pi')x'' \succ \pi' x' + (1-\pi')x''$. *Q.E.D.*

The theorem has a corollary for the case of expected utility:

COROLLARY *If the decision-maker satisfies axiom Preference Relation and maximizes expected utility and there exist* $x, x', x'' \in X'$ *and* $\pi\epsilon(0;1]$ *such that* $\pi x + (1-\pi)x'' \succcurlyeq \pi x' + (1-\pi)x''$*, then for all* $\pi'\epsilon(0;1]$ *:* $\pi' x + (1-\pi')x'' \succcurlyeq \pi' x' + (1-\pi')x''$.

The corollary holds since expected utility's independence axiom implies the axioms of FOSD and Strict FOSD. Note that in the thought experiment and experimental setup, d affects the recipient only via the payoff x_2^C. Thus, the theorem applies even to situations where the DM cares about not only the recipient's outcome but also about the recipient's opinion or feelings about the DM or her decision d. Thus, for consequentialist preferences, even allowing such consequences as others' opinions, the DM's optimal split does not depend on the probability of the DM's split being implemented.

Formally, our theorem needs both strict FOSD and weak FOSD since the former does not imply the latter.[24] Are there assumptions besides strict FOSD so we don't need both strict and weak FOSD? Yes, if a preference satisfies Preference Relation, Strict FOSD, Continuity, and Rich Domain then it satisfies FOSD.[25]

DEFINITION $\succsim$ is continuous if for all $p, q, r \in P$ the sets $\{\alpha \in [0,1] : \alpha p + (1-\alpha)q \succsim r\}$ and $\{\alpha \in [0,1] : r \succsim \alpha p + (1-\alpha)q\}$ are closed in [0,1].

Now consider someone who would like to be fair, but between two unfair lotteries she prefers the one that is more fair. Formally, for all $\pi, \pi' \in [0;1] : \pi \cdot (1-\pi) \geq \pi' \cdot (1-\pi')$ if and only if $(x; \pi, y; 1-\pi) \succsim (x; \pi', y; 1-\pi')$. The axiom of Strict FOSD is trivially satisfied since there is no lottery that strictly first-order stochastically dominates another lottery. Axiom of continuity is satisfied. However, axiom of FOSD is violated: $(x; \frac{2}{3}, y; \frac{1}{3})$ weakly first order-stochastically dominates $(x; \frac{1}{2}, y; \frac{1}{2})$, but $(x; \frac{1}{2}, y; \frac{1}{2}) \succ (x; \frac{2}{3}, y; \frac{1}{3})$.

AXIOM *(Continuity)* $\succsim$ *is continuous.*

[24] The axiom of Strict FOSD does not imply the axiom of FOSD. An example can be derived from Machina (1989) with preferences that satisfy Preference Relation and Strict FOSD but violate FOSD.

[25] Continuity alone is not sufficient for the axiom of Strict FOSD to imply the axiom of FOSD.

AXIOM *(Rich domain) There are two outcomes $x, y \in X$ such that $x \succ y$.*

PROPOSITION *If a preference satisfies Preference Relation, Strict FOSD, Continuity, and Rich Domain then it satisfies FOSD.*

PROOF: Suppose p weakly first-order stochastically dominates q. We need to show that $p \succsim q$.
Suppose not, that is $q \succ p$.

Since X is finite there exits an $\overline{x}$, $\underline{x}$ such that for all x: $\overline{x} \succsim x$, and an $x \succsim \underline{x}$. By the axiom of Rich Domain, $\overline{x} \succ \underline{x}$.

At least one of the following three cases is satisfied: (i) $\overline{x} \succ q$, (ii) $p \succ \underline{x}$ or (iii) $q \succsim \overline{x} \succ \underline{x} \succsim p$.

(i) Since p weakly first-order stochastically dominates q, and $\overline{x} \succ q$, for any $\alpha > 0$ the lottery $\alpha\overline{x} + (1-\alpha)p$ strictly first-order stochastically dominates q. But then $\{\alpha : \alpha\overline{x} + (1-\alpha)p \succsim q\} = (0, 1]$, a violation of continuity.

(ii) Since p weakly first-order stochastically dominates q, and $p \succ \underline{x}$, for any $\alpha > 0$, p strictly first-order stochastically dominates $\alpha\underline{x} + (1-\alpha)q$. But then $\{\alpha : p \succsim \alpha\underline{x} + (1-\alpha)q\} = (0, 1]$, a violation of continuity.

(iii) First we show that all elements z in the support of q satisfy $z \sim \overline{x}$. First note that by definition of $\overline{x}$, all elements in the support satisfy $\overline{x} \succsim z$. Suppose there is at least one element z such that $\overline{x} \succ z$, then $\overline{x}$ strictly first-order stochastically dominates q, which by axiom Strict FOSD implies $\overline{x} \succ q$, a contradiction. Thus, for all elements z in the support of q we have $z \sim \overline{x}$.

Second, we show that all elements z in the support of p satisfy $z \sim \underline{x}$. First note that by definition of $\underline{x}$, all elements in the support satisfy $z \succsim \underline{x}$. Suppose there is at least one element z such that $z \succ \underline{x}$, then p strictly first-order stochastically dominates $\underline{x}$, which by axiom SFOSD implies $p \succ \underline{x}$, a contradiction. Thus for all elements z in the support of p we have $z \sim \underline{x}$.

Since all elements in the support of q are indifferent to $\overline{x}$, all elements in the support of p are indifferent to $\underline{x}$, and $\overline{x}$ is strictly preferred to $\underline{x}$, q strictly first order stochastically dominates p. But that is a contradiction to p weakly first order stochastically dominating q. *Q.E.D.*

Further note that if the cardinality of the outcome space is 2, then independence is as weak an axiom as first-order stochastic dominance.

AXIOM *(Independence) $\succsim$ satisfies independence if for all lotteries p, q, r in P : $p \succcurlyeq q \Leftrightarrow \alpha p + (1-\alpha)r \succcurlyeq \alpha q + (1-\alpha)r$.*

PROPOSITION *Consider* X *with 2 elements. If $\succsim$ on P(X) satisfies Preference Relation, Strict FOSD and FOSD, then it satisfies Independence.*

PROOF: Without loss of generality let $X = \{x, y\}$ and $x \succcurlyeq y$. Denote $k = \alpha p + (1-\alpha)r$ and $l = \alpha q + (1-\alpha)r$.

(i) $x \sim y$

Then l weakly first-order stochastically dominates k, and vice versa. Thus by FOSD $l \succcurlyeq k$ and $k \succcurlyeq l$, thus $k \sim l$.

(ii) $x \succ y$

(ii.i) p and q are identical: Then $k = l$ and trivially $k \sim l$.

(ii.ii) $p \sim q$ but not identical: Then one must strictly first-order stochastically dominate the other, which by Strict FOSD contradicts indifference.

(ii.iii) $p \succ q$: By the lemma below, this implies $p(x) > q(x)$, and thus $p(y) < q(y)$, then k strictly first-order stochastically dominates l:

For y: $\sum_{y \succsim z} k(z) = k(y) = \alpha p(y) + (1-\alpha) r(y) < \alpha q(y) + (1-\alpha) r(y) = l(y) = \sum_{y \succsim z} l(z)$.
For x: $\sum_{x \succsim z} k(z) = 1 = \sum_{x \succsim z} l(z)$.
Thus by Strict FOSD $l \succ k$. *Q.E.D.*

LEMMA *Consider $X = \{x, y\}$ and $x \succ y$. If $\succsim$ on P(X) satisfies Preference Relation and Strict FOSD, then $p \succ q$ if and only if $p(x) > q(x)$.*

PROOF: 1.) $p \succ q$ implies $p(x) > q(x)$.

Proof by Contradiction: Suppose $p(x) \leq q(x)$.

i) $p(x) = q(x)$: This implies that $p = q$, and thus trivially by completeness $p \sim q$, a contradiction.

ii) $p(x) < q(x)$: Since $x \succ y$ this means that q strictly first order stochastically dominates p, and thus by Strict FOSD $q \succ p$, a contradiction.

2.) $p(x) > q(x)$ implies $p \succ q$: This follows from Strict FOSD. *Q.E.D.*

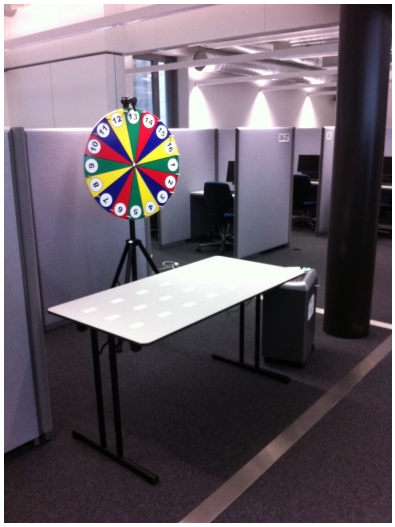
Note that there are examples where Independence is violated but FOSD is not. Cumulative prospect theory is one such example where the Allais paradox is allowed (thus violating Independence) but FOSD is satisfied.

Next, we illustrate consequentialist-deontological preferences where the optimal decision changes as the probability of being consequentialist changes. Let $u(x_1, d) = f(x_1) + b(d)$. Then, $U(x_1, d) = \pi(f(x_1^C) + b(d)) + (1-\pi)(f(x_1^N) + b(d))$ and $V(d) = \pi f(\omega - d) + (1-\pi) f(\omega - \kappa) + b(d)$. The first order condition is: $\frac{\partial V(d)}{\partial d} = -\pi f_1(\omega - d) + b_1(d) = 0$. For d^* to be a maximum, the second order condition yields: $\frac{\partial^2 V(d)}{\partial d^2} = \pi f_{11}(\omega - d) + b_{11}(d) < 0$. Applying the implicit function theorem to the first order condition yields: $\frac{\partial d^*}{\partial \pi} = \frac{f_1(\omega - d^*)}{\pi f_{11}(\omega - d^*) + b_{11}(d^*)} < 0$, since utility is increasing in its own outcomes and the denominator which is the second derivative of the indirect objective function is negative.

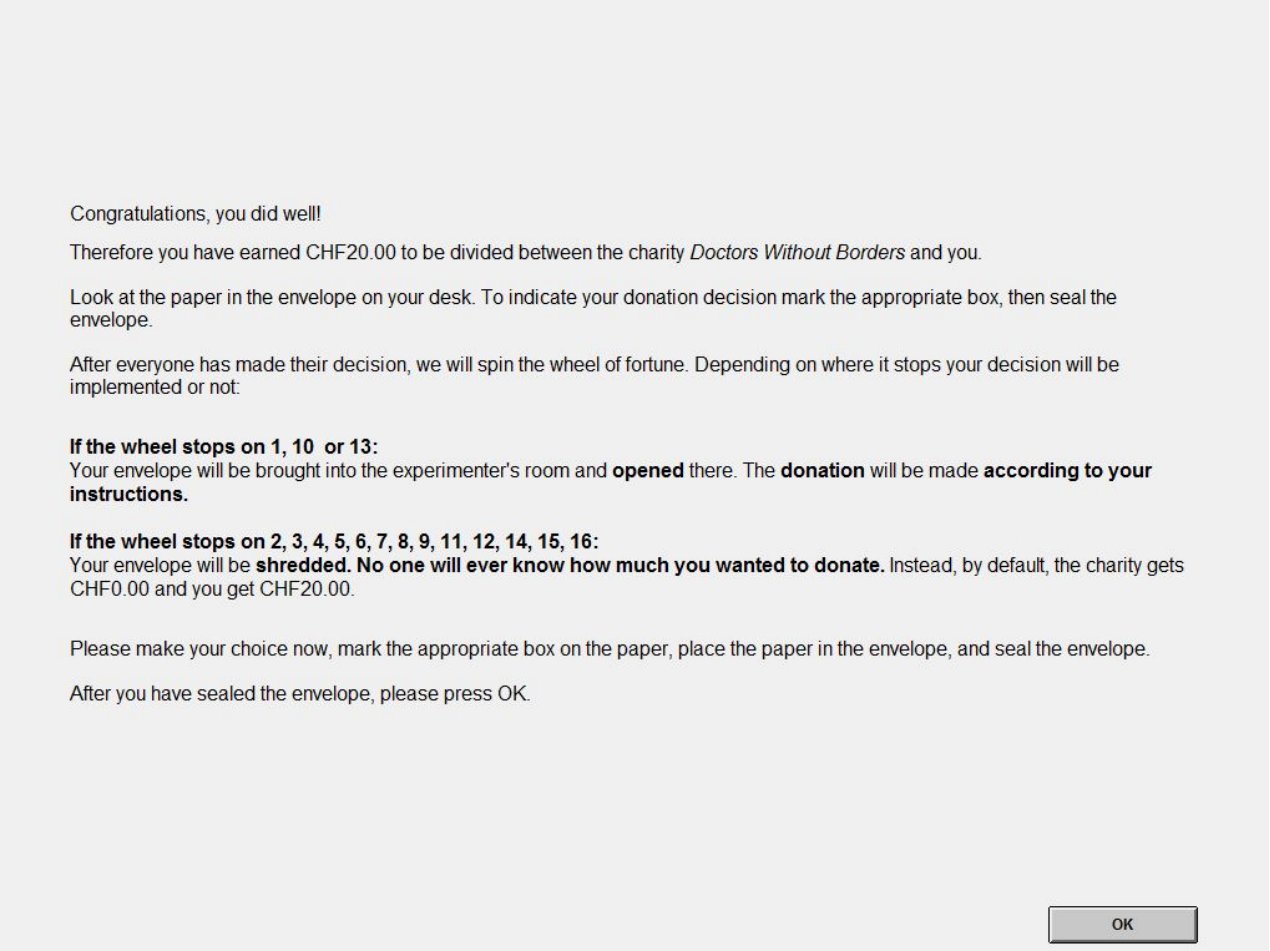
Note that the recipient's payoff is a function of the DM's payoffs, but as long as other-regarding concerns are concave then the sum of utility from DM's own payoffs and utility from others' payoffs is still concave and the above result holds. Decisions do not have to be continuous to obtain this result. If decisions are discrete, then the behavior of a mixed consequentialist-deontological person is jumpy (i.e., it weakly increases as her decision becomes less consequential).

APPENDIX B: LAB INSTRUCTIONS

Appendix Figure 1.— Lab Implementation

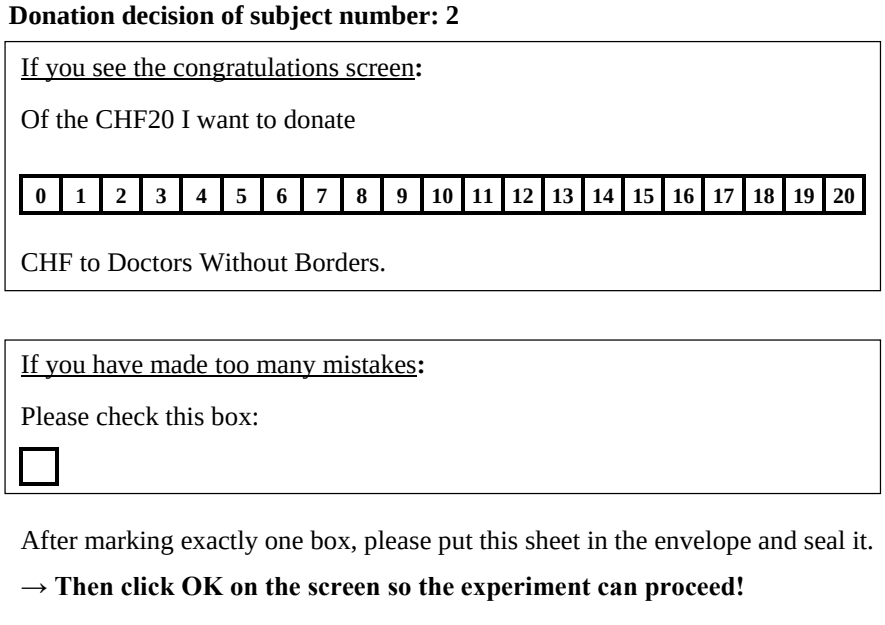


Appendix Figure 2.— Sample Screenshot of Lab Experiment



Notes: Shredding Experiment Instructions Donation Screen for Subject with $\pi = 3/16$ and $\kappa = 0$.

APPENDIX FIGURE 3.— Donation Decision Placed in Sealed Envelope



Donation decision of subject number: 2

If you see the congratulations screen**:**

Of the CHF20 I want to donate

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

CHF to Doctors Without Borders.

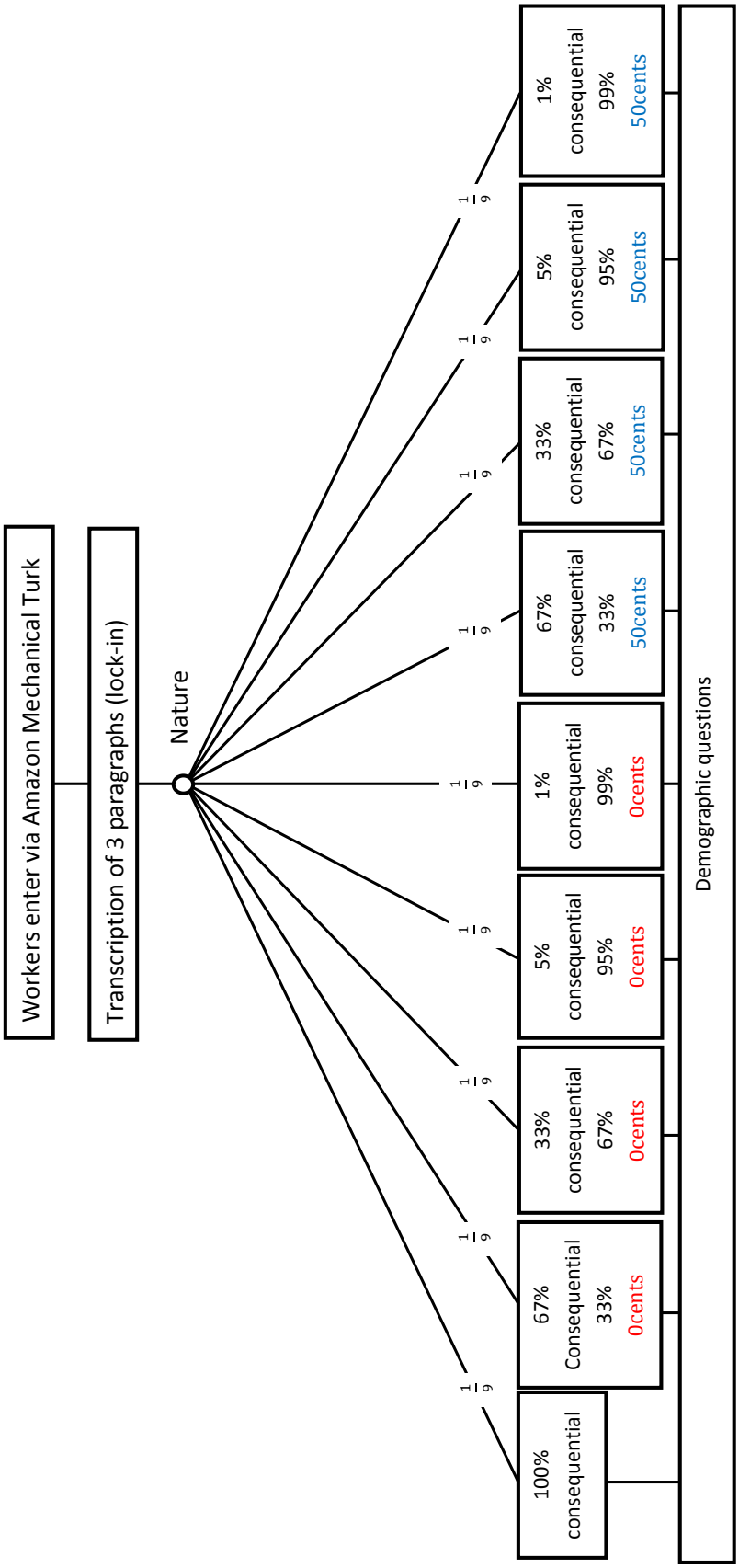
If you have made too many mistakes**:**

Please check this box: ☐

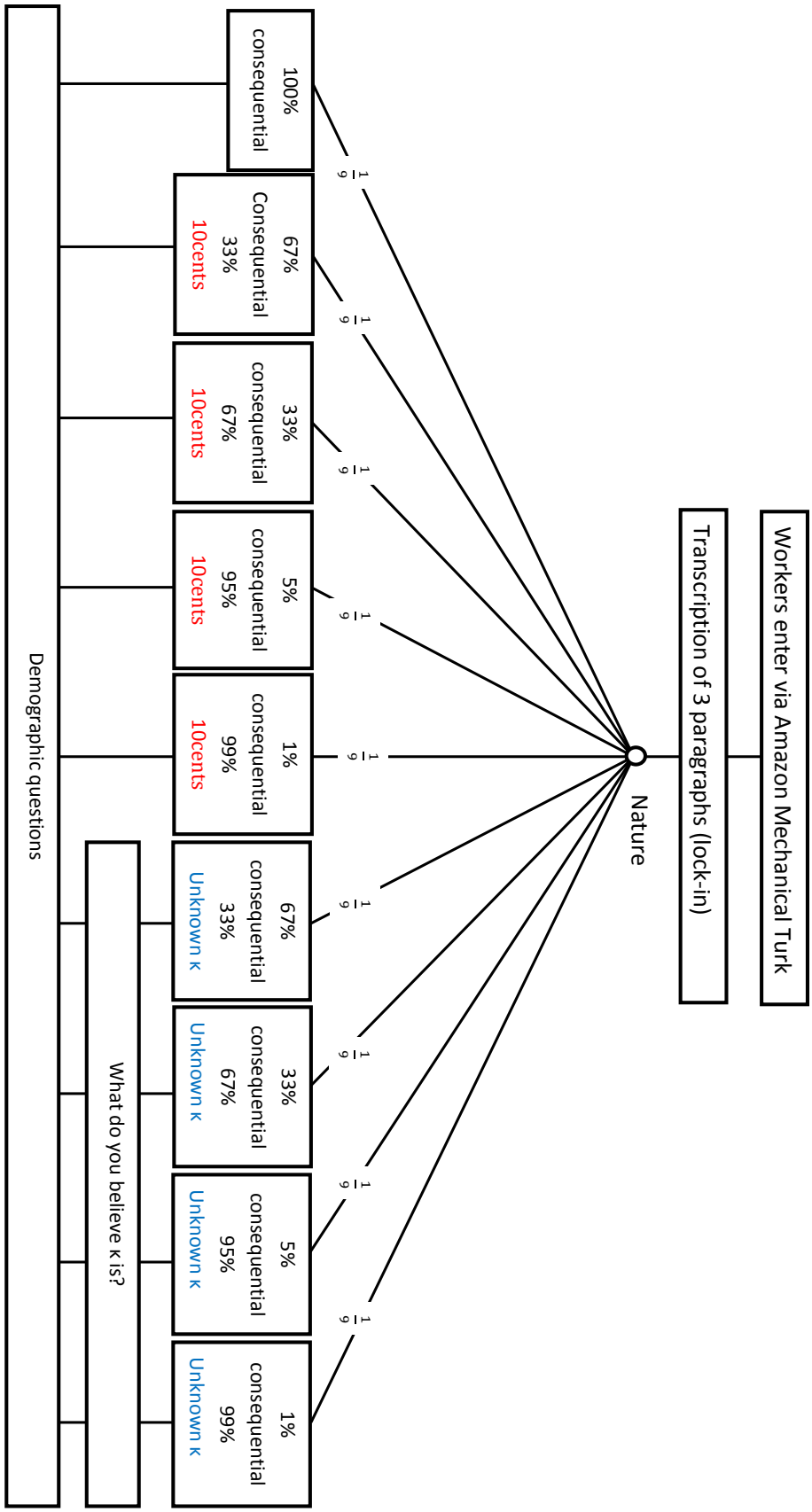
After marking exactly one box, please put this sheet in the envelope and seal it.

→ Then click OK on the screen so the experiment can proceed!

APPENDIX C: MTURK INSTRUCTIONS



APPENDIX FIGURE 4.— Schematic of MTurk Experiment (Experiment 1)



APPENDIX FIGURE 5.— Schematic of MTurk Experiment (Experiment 2)